auxilia na atribuição de probabilidades. Assim, para a variável aleatória contínua X representando a profundidade do lençol de água, a função densidade f é dada por

$$f(x) = \begin{cases} 1/80, & \text{para } 20 \leq x \leq 100; \\ 0, & \text{para } x < 20 \text{ ou } x > 100. \end{cases}$$

Tendo em vista que, nesse exemplo, a função densidade é bastante simples, a probabilidade de que a profundidade do lençol esteja em um dado intervalo pode ser calculada com o uso de áreas de figuras planas. Assim, para obter a probabilidade de uma profundidade pelo menos igual a 25, mas inferior a 29, calculamos a área do retângulo:

, portanto, $P(25 \leq X < 29) = 4/80$. □

Considerando o caso geral, vamos nos ocupar agora em formalizar as déias discutidas anteriormente. Faremos isso através da definição apresentada a eguir.

Definição 6.1: Função densidade de probabilidade

Dizemos que $f(x)$ é uma função contínua de probabilidade ou função ensidade de probabilidade para uma variável aleatória contínua X, se satisfaz uas condições:

i) $f(x) \geq 0$, para todo $x \in (-\infty, \infty)$;

ii) A área definida por $f(x)$ é igual a 1.

□

Com o auxílio do cálculo diferencial e integral, podemos caracterizar a condição ii) através de

$$\int_{-\infty}^{\infty} f(x)\, dx = 1\,.$$

Da mesma forma, para calcular probabilidades, temos que para $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx\,;$$

a integral, acima, indica a área sob a função f definida pelo intervalo $[a, b]$.

Note que, pela forma como atribuímos as probabilidades no caso contínuo, teremos área zero sob qualquer valor individual, isto é, $P(X = k) = 0$ para qualquer k. Portanto, em se tratando de variáveis aleatórias contínuas, a probabilidade de ocorrência de um valor isolado é sempre zero e, consequentemente, as probabilidades calculadas sobre os intervalos $[a, b]$, $[a, b)$, $(a, b]$ e (a, b) são as mesmas, para quaisquer valores de a e b.

Exemplo 6.2: Arqueólogos estudaram uma certa região e estabeleceram um modelo teórico para a variável C, *comprimento de fósseis da região* (em cm). Suponha que C é uma variável aleatória contínua com a seguinte função densidade de probabilidade:

$$f(c) = \begin{cases} \frac{1}{40}(\frac{c}{10} + 1) & \text{se } 0 \leq c \leq 20\,; \\ 0 & \text{caso contrário.} \end{cases}$$

É imediato observar que $f(c)$ é positiva. Através do gráfico da função, apresentado a seguir, podemos verificar com auxílio da fórmula da área de trapézio que

$$\text{área sob } f(c) = \frac{\frac{1}{40} + \frac{3}{40}}{2} \times 20 = 1.$$

Concluímos que $f(c)$ é efetivamente uma densidade. Tendo em vista a forma simples de $f(c)$, o cálculo de probabilidades de interesse para esse exemplo poderá ser feito sem dificuldades através de áreas.

A probabilidade de um fóssil, escolhido ao acaso nessa região, apresentar comprimento inferior a 8 cm pode ser calculada diretamente do gráfico da função densidade de probabilidade, como mostra a figura a seguir:

O cálculo da probabilidade envolve a soma de duas áreas:

e, assim, temos que $P(C < 8) = 7/25$. □

Exemplo 6.3: Num teste educacional com crianças, o tempo para a realização de uma bateria de questões de raciocínio verbal e lógico é medido e anotado para ser comparado com um modelo teórico. Este teste é utilizado para identificar o desenvolvimento das crianças e auxiliar a aplicação de medidas corretivas. O modelo teórico considera T, *tempo de teste em minutos*, como uma variável aleatória contínua com função densidade de probabilidade dada por:

$$f(t) = \begin{cases} \frac{1}{40}(t-4), & \text{se } 8 \leq t < 10; \\ \frac{3}{20}, & \text{se } 10 \leq t \leq 15; \\ 0, & \text{caso contrário.} \end{cases}$$

O gráfico da função densidade é apresentado a seguir. Deve ser notado que, pela definição de $f(t)$, ela se anula para $t < 8$ ou $t > 15$.

Solicitamos ao leitor que verifique que a função $f(t)$ satisfaz a definição de densidade. Para calcular $P(9 < T \leq 12)$, vamos obter a área sob $f(t)$ no intervalo $(9, 12]$:

Segue, sem maiores dificuldades, que $P(9 < T \leq 12) = 7/16$, valor esse obtido pela soma do trapézio definido no intervalo $(9, 10)$ com o retângulo determinado pelo intervalo $[10, 12]$ (veja a figura).

Através do uso de integral, essa mesma probabilidade seria calculada da seguinte forma:

$$\begin{aligned} P(9 < T \leq 12) &= \int_9^{12} f(t)\,dt \\ &= \int_9^{10} f(t)\,dt + \int_{10}^{12} f(t)\,dt \\ &= \int_9^{10} \frac{1}{40}(t-4)\,dt + \int_{10}^{12} \frac{3}{20}\,dt \\ &= \frac{1}{40}\left(\frac{t^2}{2} - 4t\right)\Big|_9^{10} + \frac{3}{20}\,t\Big|_{10}^{12} \\ &= \frac{11}{80} + \frac{6}{20} \\ &= \frac{7}{16}. \end{aligned}$$

A aplicação da integral foi dividida em duas partes, pois a função $f(t)$ é diferente nos intervalos $(9, 10)$ e $[10, 12]$. □

Vamos, agora, apresentar as expressões para valor esperado, mediana, moda e variância no caso contínuo. A interpretação de cada uma dessas grandezas é semelhante àquela discutida no caso discreto. Algumas das expressões são alteradas devido à nova forma de atribuição de probabilidades.

Definição 6.2: Medidas de posição para variáveis aleatórias contínuas

O valor esperado ou média da variável aleatória contínua X, com função densidade dada por $f(x)$, é dada pela expressão:

$$E(X) = \mu = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

A mediana é o valor Md que tem a propriedade de:

$$P(X \geq Md) \geq 0{,}5 \text{ e } P(X \leq Md) \geq 0{,}5.$$

A moda é o valor Mo tal que,

$$f(Mo) = \max_x f(x).$$

□

Observe que a definição de mediana é idêntica ao caso discreto. A média teve sua expressão alterada com a substituição da somatória pela integral e de p_i por $f(x)\,dx$. Para a moda, precisamos tomar o máximo da função densidade e, como antes, ela não é necessariamente única. A notação para o caso contínuo será a mesma utilizada para as variáveis aleatórias discretas.

Definição 6.3: Variância para variáveis aleatórias contínuas

Para uma variável aleatória X com densidade $f(x)$, a variância é dada por

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx.$$

□

Como no caso discreto, a variância é a medida de dispersão mais utilizada na prática. Aqui podemos, também, utilizar a expressão alternativa

$$\sigma^2 = E(X^2) - \mu^2,$$

com $E(X^2)$ sendo calculado como:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx\,.$$

O desvio padrão é a raiz quadrada da variância e, como já mencionado anteriormente, tem a mesma unidade de medida da variável original, o que facilita a interpretação dos seus valores.

Exemplo 6.4: A variável C apresentada no Exemplo 6.2 tinha a seguinte função densidade:

$$f(c) = \begin{cases} \frac{1}{40}(\frac{c}{10}+1), & \text{se } 0 \le c \le 20\,; \\ 0\,, & \text{caso contrário.} \end{cases}$$

Vamos determinar a média e a variância de C. Temos,

$$\mu = \int_0^{20} c\,\frac{1}{40}\,(\frac{c}{10}+1)\,dc = \frac{1}{400}\frac{c^3}{3}\Big|_0^{20} + \frac{1}{40}\frac{c^2}{2}\Big|_0^{20} =$$
$$= \frac{20}{3} + 5 = \frac{35}{3}\,.$$

Para a variância, calculamos primeiro $E(C^2)$:

$$E(C^2) = \int_0^{20} c^2\,\frac{1}{40}\,(\frac{c}{10}+1)\,dx = \frac{1}{400}\frac{c^4}{4}\Big|_0^{20} + \frac{1}{40}\frac{c^3}{3}\Big|_0^{20} =$$
$$= 100 + \frac{200}{3} = \frac{500}{3}\,.$$

Então,

$$\sigma_C^2 = E(C^2) - \mu^2 = \frac{500}{3} - \left(\frac{35}{3}\right)^2 = \frac{275}{9} = 30{,}56\ \text{cm}^2.$$

Logo, o desvio padrão é $\sigma_C = \sqrt{30{,}56} = 5{,}53$ cm.

A moda segue diretamente do gráfico da função densidade e é igual a 20. Para obter a mediana, notamos inicialmente que a densidade é uma função contínua com mesma expressão no intervalo entre 0 e 20. Dessa forma, para atender a condição da definição de mediana, basta resolver a equação:

$$P(C \ge Md) = 0{,}5.$$

Assim,

$$\int_{Md}^{20} \frac{1}{40}\,(\frac{c}{10}+1)\,dc = \frac{1}{400}\frac{c^2}{2}\Big|_{Md}^{20} + \frac{1}{40}c\Big|_{Md}^{20} = 0{,}5.$$

Resultando na equação do 2o. grau:

$$Md^2 + 20Md - 400 = 0,$$

cuja solução é $Md = 12{,}36$ (o outro valor é abandonado por ser negativo). □

As propriedades do valor esperado e da variância apresentadas para variáveis aleatórias discretas permanecem válidas e a verificação pode ser feita através das propriedades da integral. A distribuição conjunta de duas ou mais variáveis aleatórias contínuas é definida através da função densidade conjunta de probabilidade. As idéias básicas são as mesmas do caso discreto, porém requerem um melhor conhecimento de cálculo diferencial e integral, envolvendo integrais duplas. Não desenvolveremos esse tópico e recomendamos ao leitor interessado a consulta às referências.

Exercícios da Seção 6.1:

1. Verifique se as expressões a seguir são funções densidade de probabilidade (assuma que elas se anulam fora dos intervalos especificados).

a. $f(x) = 3x$, se $0 \le x \le 1$.

b. $f(x) = x^2/2$, $x \ge 0$.

c. $f(x) = (x-3)/2$, $3 \le x \le 5$.

d. $f(x) = 2$, $0 \le x \le 2$.

e. $f(x) = \begin{cases} (2+x)/4, & \text{se } -2 \le x < 0; \\ (2-x)/4, & \text{se } 0 \le x \le 2. \end{cases}$

f. $f(x) = -\pi$, se $-\pi < x < 0$.

2. O tempo, em minutos, de digitação de um texto por secretárias experientes é uma variável aleatória contínua X. Sua densidade é apresentada a seguir.

$$f(x) = \begin{cases} 1/4, & \text{se } 0 \le x < 2; \\ 1/8, & \text{se } 2 \le x < 6; \\ 0, & \text{caso contrário.} \end{cases}$$

Determine:

a. $P(X > 3)$.

b. $P(1 < X \leq 4)$.

c. $P(X < 3 | X \geq 1)$.

d. Um número b tal que $P(X > b) = 0{,}6$.

e. O valor esperado, a variância e a moda de X.

3. A quantia gasta anualmente, em milhões de reais, na manutenção do asfalto em uma cidade do interior é representada pela variável Y com densidade dada por:

$$f(y) = \begin{cases} \frac{8}{9}y - \frac{4}{9}, & \text{se } 0{,}5 \leq y < 2; \\ 0, & \text{caso contrário.} \end{cases}$$

Obtenha:

a. $P(Y < 0{,}8)$.

b. $P(Y > 1{,}5 | Y \geq 1)$.

c. O valor esperado e a variância de Y.

d. A mediana de Y.

4. O gráfico abaixo representa a densidade de uma variável aleatória X.

a. Obtenha o valor de a.

b. Determine $P(X > 0 | X < 3)$.

c. Calcule $Md(X)$, $E(X)$ e $Var(X)$.

5. Numa certa região, fósseis de pequenos animais são freqüentemente encontrados e um arqueólogo estabeleceu o seguinte modelo de probabilidade para o comprimento, em centímetros, desses fósseis.

$$f(x) = \begin{cases} \frac{1}{40}x, & 4 \leq x \leq 8; \\ \frac{-1}{20}x + \frac{3}{5}, & 8 \leq x \leq 10; \\ \frac{1}{10}, & 10 \leq x \leq 11; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Faça um gráfico da função densidade.

b. Para um fóssil encontrado nessa região, determine a probabilidade do comprimento ser inferior a 6 centímetros? E de ser superior a 5 mas inferior a 10,5 cm?

c. Encontre o valor esperado para o comprimento dos fósseis da região.

6.2 Principais Modelos Contínuos

Apresentamos, nesta seção, os principais modelos teóricos para variáveis aleatórias contínuas. Vimos que, para caracterizar completamente uma variável aleatória contínua, precisamos fornecer sua função densidade de probabilidade que, segundo sua definição, é uma função positiva e com integral igual a 1.

Definição 6.4: Modelo Uniforme Contínuo

Uma variável aleatória X tem distribuição *Uniforme Contínua* no intervalo $[a, b]$, $a < b$, se sua função densidade de probabilidade é dada por:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b; \\ 0, & \text{caso contrário.} \end{cases}$$

Usaremos a notação $X \sim U[a, b]$ para indicar que X segue o modelo Uniforme Contínuo no intervalo considerado.

Note que não há restrição de valores para a e b, exceto o fato de $a < b$. A Figura 6.1 mostra a densidade do modelo $U[a, b]$, para $a, b > 0$. □

Figura 6.1: Densidade Uniforme Contínua.

O modelo Uniforme pressupõe que os valores possíveis para a variável aleatória têm todos a mesma probabilidade de ocorrência. Seu valor esperado e sua variância são obtidos através do cálculo de integrais, de tal forma que:

$$\mu = \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{b-a} \frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2};$$

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{1}{b-a} \frac{x^3}{3}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3};$$

logo,

$$\sigma^2 = E(X^2) - \mu^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Exemplo 6.5: Com o objetivo de verificar a resistência à pressão de água, os técnicos de qualidade de uma empresa inspecionam os tubos de PVC produzidos. Os tubos inspecionados têm 6 metros de comprimento e são submetidos a grandes pressões até o aparecimento do primeiro vazamento, cuja distância a uma das extremidades (fixada à priori) é anotada para fins de análise posterior. Escolhe-se um tubo ao acaso para ser inspecionado. Queremos calcular a probabilidade de que o vazamento esteja, no máximo, a 1 metro das extremidades.

Vamos denotar por X a variável aleatória que indica a distância correspondente ao vazamento. Admitindo igual probabilidade de ocorrência em

todos os pontos, temos que $X \sim U[0,6]$, com função densidade de probabilidade dada por

$$f(x) = \begin{cases} 1/6, \text{ se } 0 \le x \le 6; \\ 0, \text{ caso contrário.} \end{cases}$$

Para calcular a probabilidade de $X \in \{[0,1] \cup [5,6]\}$, podemos obter as áreas dos dois retângulos hachuriados na figura a seguir.

Segue, sem maiores dificuldades, que a probabilidade desejada é 1/3.

Esse mesmo cálculo poderia ser feito através de integrais da seguinte forma:

$$\begin{aligned} P(X \in \{[0,1] \cup [5,6]\}) &= P(0 \le X \le 1) + P(5 \le X \le 6) \\ &= \int_0^1 \frac{1}{6} dx + \int_5^6 \frac{1}{6} dx \\ &= \frac{x}{6}\Big|_0^1 + \frac{x}{6}\Big|_5^6 \\ &= \frac{1}{6} - 0 + \frac{6}{6} - \frac{5}{6} = \frac{1}{3}. \end{aligned}$$

Note que os intervalos $[0,1]$ e $[5,6]$ são disjuntos e, portanto, a probabilidade da sua união é a soma das probabilidades de ocorrência de cada intervalo. □

Definição 6.5: Modelo Exponencial

Uma variável aleatória contínua X, assumindo valores não negativos, segue o modelo *Exponencial* com parâmetro $\alpha > 0$ se sua densidade é

$$f(x) = \begin{cases} \alpha\, e^{-\alpha x}, & x \geq 0; \\ 0, & \text{caso contrário.} \end{cases}$$

A densidade está representada graficamente na Figura 6.2 e adotaremos a notação $X \sim Exp(\alpha)$ para indicar que X tem distribuição Exponencial de parâmetro α. □

Figura 6.2: Densidade Exponencial.

Para calcular probabilidades com a Exponencial, precisamos resolver a integral correspondente, já que não teremos as figuras geométricas simples dos exemplos considerados até aqui. Assim,

$$P(a < X < b) = \int_a^b \alpha\, e^{-\alpha x} dx = -\, e^{-\alpha x}\Big|_a^b = e^{-\alpha a} - e^{-\alpha b}.$$

Note que a inclusão ou não dos extremos a e b não altera o cálculo efetuado acima.

Para obter a média e a variância, será necessário aplicar a integração por partes, porém, não vamos fazer esse cálculo e, apenas, apresentamos as

expressões finais. Temos, para $X \sim Exp(\alpha)$,

$$\mu = 1/\alpha \quad e \quad \sigma^2 = 1/\alpha^2.$$

Exemplo 6.6: Uma indústria fabrica lâmpadas especiais que ficam em operação continuamente. A empresa oferece a seus clientes a garantia de reposição, caso a lâmpada dure menos de 50 horas. A vida útil dessas lâmpadas é modelada através da distribuição Exponencial com parâmetro 1/8000. Determine a proporção de trocas por defeito de fabricação.

Cada lâmpada terá seu tempo de duração definido pela particular realização da variável aleatória. Isto é, a vida útil da lâmpada pode ser pensada como um valor escolhido de acordo com a densidade Exponencial de parâmetro 1/8000. Representemos, pela variável aleatória T, o *tempo de vida da lâmpada*, e assim $T \sim Exp(1/8000)$. A probabilidade desejada será

$$P(T < 50) = \int_0^{50} \frac{1}{8000} e^{-\frac{1}{8000}t}\, dt = 1 - e^{-\frac{50}{8000}} = 0{,}006.$$

Dessa forma, a proporção de trocas por defeito de fabricação será de aproximadamente 0,6%. Esse número é relativamente pequeno, o que não é surpresa, tendo em vista que, como o parâmetro é $\alpha = 1/8000$, a duração média das lâmpadas é $\mu = 1/\alpha = 8000$ horas. □

A distribuição Exponencial tem sido amplamente utilizada nas áreas de física, engenharia, computação e biologia, entre outras. Variáveis como a vida útil de equipamentos, tempos de falha, tempos de sobrevivência de espécies e intervalos entre solicitações de recursos são algumas das quantidades que têm sido modeladas, com bons resultados, pela Exponencial. Essa densidade tem, ainda, a vantagem de ter propriedades matemáticas interessantes, conforme veremos no próximo exemplo.

Exemplo 6.7: O intervalo de tempo, em minutos, entre emissões consecutivas de uma fonte radioativa é uma variável aleatória com distribuição Exponencial de parâmetro $\alpha = 0{,}2$. Vamos calcular a probabilidade de haver uma emissão em um intervalo inferior a 2 minutos. Temos,

$$P(X < 2) = \int_0^2 0{,}2\, e^{-0{,}2x}\, dx = -\, e^{-0{,}2x}\Big|_0^2 = -\, e^{-0{,}4} + 1 = 0{,}33.$$

Calculemos, agora, a probabilidade do intervalo ser superior ou igual a 7, sabendo-se que ele é superior ou igual a 5 minutos.

$$P(X \geq 7 | X \geq 5) = \frac{P(X \geq 7, X \geq 5)}{P(X \geq 5)} = \frac{P(X \geq 7)}{P(X \geq 5)} =$$

$$= \frac{\int_7^\infty 0{,}2\, e^{-0{,}2x}\, dx}{\int_5^\infty 0{,}2\, e^{-0{,}2x}\, dx} = \frac{e^{-1{,}4}}{e^{-1}} = 0{,}67\,.$$

Por outro lado, $P(X \geq 2)$ pode ser calculada pelo complementar de $P(X < 2)$, resultando em 0,67. Note que a igualdade desses valores, sugere que:

$$P(X \geq 7 | X \geq 5) = P(X \geq 2).$$

Como veremos, adiante, isto não foi coincidência mas sim uma importante propriedade da densidade Exponencial. Em palavras, a informação de que o intervalo é maior ou igual a 5, faz com que a probabilidade dele ser maior ou igual a 7 possa ser calculada através de uma translação no tempo. Isto é, podemos assumir que a origem do tempo é 5 e, portanto, a diferença $7 - 5 = 2$ seria o tempo que devemos considerar para calcular a probabilidade desejada. □

A característica de permitir a translação da origem no cálculo de probabilidades é uma propriedade bastante importante da Exponencial e é conhecida como *falta de memória*. Ela é a única distribuição contínua com essa propriedade, conforme verificamos a seguir. Para tal, seja $X \sim Exp(\alpha)$ e $s, t > 0$, temos que

$$P(X \geq t+s | X \geq s) = \frac{P(X \geq t+s, X \geq s)}{P(X \geq s)} = \frac{P(X \geq t+s)}{P(X \geq s)} =$$

$$= \frac{\int_{t+s}^\infty \alpha\, e^{-\alpha x} dx}{\int_s^\infty \alpha\, e^{-\alpha x} dx} = \frac{-\, e^{-\alpha x}|_{t+s}^\infty}{-\, e^{-\alpha x}|_s^\infty} = \frac{e^{-\alpha(t+s)}}{e^{-\alpha s}}$$

$$= e^{-\alpha t} = P(X \geq t)\,.$$

Supondo que X representa o tempo de vida de um equipamento, podemos fazer a seguinte interpretação para a propriedade da falta de memória: a probabilidade do equipamento durar pelo menos $t + s$ anos, sabendo-se que já durou s, é igual à probabilidade de um equipamento novo durar pelo menos t anos. Em outras palavras, a informação da "idade" do equipamento pode ser esquecida e o que importa, para o cálculo da probabilidade, é quantos anos a mais queremos que dure.

Dentre todos os modelos teóricos, sejam contínuos ou discretos, o mais importante é o modelo Normal. Ele é muito utilizado em aplicações e também serve como aproximação para muitas outras distribuições.

Definição 6.6: Modelo Normal

Dizemos que uma variável aleatória contínua X tem distribuição *Normal* com parâmetros μ e σ^2, se sua função densidade é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,, \text{ para } -\infty < x < \infty\,.$$

Usaremos a notação $X \sim N(\mu, \sigma^2)$, para indicar que X tem distribuição Normal com parâmetros μ e σ^2. A densidade é representada na Figura 6.3. □

Algumas propriedades da densidade da Normal podem ser, facilmente, observadas de seu gráfico:

i) $f(x)$ é simétrica em relação à μ;
ii) $f(x) \to 0$ quando $x \to \pm\infty$;
iii) o valor máximo de $f(x)$ se dá para $x = \mu$.

Figura 6.3: Densidade Normal.

Pode-se ainda verificar que os parâmetros μ e σ^2 representam, respectivamente, a média e a variância da distribuição. A demonstração requer

algumas manipulações de integral e o leitor interessado pode consultar as referências no final do texto. Assim, quando indicamos que $X \sim N(\mu, \sigma^2)$, segue imediatamente que $E(X) = \mu$ e $Var(X) = \sigma^2$.

No cálculo de probabilidades para variáveis contínuas, devemos resolver a integral da função densidade no intervalo de interesse, isto é,

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx .$$

Entretanto, a integral acima só pode ser resolvida de modo aproximado e por métodos numéricos. Por essa razão as probabilidades para o modelo Normal são calculadas com o auxílio de tabelas. Para se evitar a multiplicação desnecessária de tabelas para cada par de valores (μ, σ^2), utiliza-se uma transformação que conduz sempre ao cálculo de probabilidades com uma variável de parâmetros $(0, 1)$, isto é, média 0 e variância 1.

Considere $X \sim N(\mu, \sigma^2)$ e defina uma nova variável $Z = \frac{X-\mu}{\sigma}$. Pelas propriedades do valor esperado e da variância, segue que

$$E(Z) = E(\frac{X-\mu}{\sigma}) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}[E(X)-\mu] = 0;$$

$$Var(Z) = Var(\frac{X-\mu}{\sigma}) = \frac{1}{\sigma^2}Var(X-\mu) = \frac{1}{\sigma^2}Var(X) = 1.$$

Pode-se ainda verificar que essa transformação não afeta a normalidade e, assim, a variável aleatória Z terá distribuição $N(0,1)$ e será denominada de *Normal Padrão* ou *Normal Reduzida*. Para determinar a probabilidade de $X \in [a, b]$, procedemos da seguinte forma:

$$\begin{aligned} P(a \leq X \leq b) &= P(a-\mu \leq X-\mu \leq b-\mu) \\ &= P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) \\ &= P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}), \end{aligned}$$

e, portanto, quaisquer que sejam os valores de μ e σ, utilizamos a Normal Padrão para obter probabilidades com a distribuição Normal.

Os valores para $P(0 \leq Z \leq z)$, $z \geq 0$ são apresentados em tabela no Apêndice A. Com a simetria da densidade Normal podemos calcular valores de probabilidades em outros intervalos. Note que a simetria também implica que a probabilidade de estar acima (ou abaixo) de zero é 0,5. Como probabilidade é sempre um número entre 0 e 1, a tabela contém apenas a parte decimal.

Por exemplo, para $X \sim N(2,9)$, temos

$$P(2 < X < 5) = P(\frac{2-2}{\sqrt{9}} < \frac{X-2}{\sqrt{9}} < \frac{5-2}{\sqrt{9}}) = P(0 < Z < 1) = 0{,}3413;$$

correspondendo à área sombreada no gráfico:

Para obter $P(0 \leq X < 2)$, usamos a simetria da Normal (ver figura a seguir):

$$\begin{aligned} P(0 \leq X < 2) &= P(\frac{0-2}{3} \leq Z < \frac{2-2}{3}) = P(-2/3 \leq Z < 0) \\ &= P(0 < Z \leq 2/3) = 0{,}2486. \end{aligned}$$

Podemos ainda calcular as probabilidades de intervalos com extremos negativos, utilizando os correspondentes intervalos na parte positiva. Um outro recurso importante no uso da tabela é a utilização do complementar. Por exemplo,

$$P(X > 3) = P(\frac{X-\mu}{\sigma} > \frac{3-2}{3}) = P(Z > 1/3)$$

$$= 0{,}5 - P(0 \leq Z < \frac{1}{3}) = 0{,}5 - 0{,}1293 = 0{,}3707.$$

A tabela também pode ser utilizada no sentido inverso, isto é, dado uma certa probabilidade, desejamos obter o valor que a originou. Por exemplo, quanto vale c tal que $P(0 < Z < c) = 0{,}4$? Procurando, no corpo da tabela, a probabilidade que mais se aproxima de 0,4 é 0,3997; correspondendo a 1,28 que será o valor de c.

Suponha, agora, que queremos encontrar d tal que $P(Z > d) = 0{,}8$. Observamos que d precisa ser negativo, pois a probabilidade desejada é maior que 1/2, que é o valor de $P(Z > 0)$. Assim, o intervalo $(d, 0)$ precisa ter probabilidade 0,3. Pela simetria da Normal, o intervalo $(0, -d)$ também tem probabilidade 0,3 . Da tabela, segue que $-d = 0{,}84$ e portanto $d = -0{,}84$.

Exemplo 6.8: Doentes, sofrendo de certa moléstia, são submetidos a um tratamento intensivo cujo tempo de cura foi modelado por uma densidade Normal, de média 15 e desvio padrão 2 (em dias).

Seja X o *tempo de cura* e, portanto, temos $X \sim N(15{,}4)$.

Caso desejemos saber que proporção desses pacientes demora mais de 17 dias para se recuperar, calculamos

$$P(X > 17) = P\Big(\frac{X-15}{\sqrt{4}} > \frac{17-15}{\sqrt{4}}\Big) = P(Z > 1) = 0{,}1587.$$

A probabilidade de um paciente, escolhido ao acaso, apresentar tempo de cura inferior a 20 dias será dada por

$$P(X < 20) = P\Big(\frac{X-15}{\sqrt{4}} < \frac{20-15}{\sqrt{4}}\Big) = P(Z < 2{,}5) = 0{,}9938.$$

Uma questão interessante seria saber o tempo máximo necessário para a recuperação de 25% dos pacientes. A proporção de cura para o conjunto de pacientes é interpretada como a probabilidade para um único paciente, genericamente escolhido. Assim, precisamos obter um valor t tal que:

$$P(X < t) = 0{,}25.$$

Então,

$$P\Big(\frac{X-15}{\sqrt{4}} < \frac{t-15}{\sqrt{4}}\Big) = P(Z < \frac{t-15}{\sqrt{4}}) = 0{,}25.$$

Com o uso da tabela (e alguma reflexão) obtemos

$$\frac{t-15}{\sqrt{4}} = -0{,}67 \Rightarrow t = 13{,}66.$$

Concluímos, então, informando que 25% dos pacientes ficarão curados antes de, aproximadamente, 14 dias.

Considere agora que 100 pacientes são escolhidos ao acaso, qual seria o número esperado de doentes curados em menos de 11 dias? Obtemos, inicialmente, a probabilidade de um paciente genérico ser curado em menos de 11 dias. Em seguida, essa probabilidade é interpretada como proporção de pacientes com cura em menos de 11 dias e é multiplicada por 100 para obter a resposta. Temos,

$$P(X < 11) = P\Big(\frac{X-15}{\sqrt{4}} < \frac{11-15}{\sqrt{4}}\Big) = P(Z < -2) = 0{,}0228.$$

Então, para 100 pacientes, o número esperado com tempo de cura inferior a 11 dias será de $100 \times 0{,}0228 \simeq 2$ pacientes. □

Na Tabela 6.1 apresentamos as expressões da média e da variância para os modelos contínuos estudados até aqui.

Tabela 6.1: Modelos contínuos - valor esperado e variância.

Variável Contínua	Valor Esperado	Variância
Uniforme $[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponencial (α)	$1/\alpha$	$1/\alpha^2$
Normal (μ, σ^2)	μ	σ^2

A distribuição Normal é uma das mais importantes em Estatística. Muitos dos fenômenos aleatórios de interesse comportam-se próximos a essa distribuição, com valores muito freqüentes em torno da média, diminuindo a freqüência à medida que nos afastamos da média. Uma outra razão da importância da Normal se refere à sua utilização como aproximação para outras distribuições. No próximo exemplo, veremos como utilizá-la para aproximar o modelo Binomial.

Exemplo 6.9: Estudo do Sindicato dos Bancários indica que cerca de 30% dos funcionários de banco têm problemas de estresse, provenientes das condições de trabalho. Numa amostra de 200 bancários, qual seria a probabilidade de pelo menos 50 com essa doença?

Admitindo que esses bancários foram sorteados ao acaso dentre todos os trabalhadores dessa categoria, podemos considerar que cada um deles tem a mesma probabilidade de ter ou não problema de estresse. Assim, o modelo Binomial poderá ser usado para a variável aleatória que conta o número total de bancários com problemas. Sendo X essa variável, temos $X \sim b(200; 0{,}3)$ e a probabilidade desejada será

$$P(X \geq 50) = \sum_{k=50}^{200} \binom{200}{k} 0{,}3^k \, 0{,}7^{200-k}.$$

Mesmo com o auxílio de uma calculadora, a obtenção do resultado será trabalhoso. Se tivermos um computador à disposição certamente a tarefa ficará mais fácil, principalmente se pudermos acessar algum *software* estatístico. Veremos como podemos calcular a probabilidade acima de modo rápido, porém aproximado, através da distribuição Normal.

Ao estudar a Binomial, vimos que se $X \sim b(200; 0{,}3)$ então $E(X) = np = 60$ e $Var(X) = np(1-p) = 42$. A idéia é considerar uma variável aleatória Normal com média e variância iguais as da Binomial que desejamos aproximar. Isto é, utilizamos $Y \sim N(\mu = 60, \sigma^2 = 42)$, de modo que:

$$P(X \geq 50) \simeq P(Y \geq 50) = P\left(\frac{Y-60}{\sqrt{42}} \geq \frac{50-60}{\sqrt{42}}\right) = P(Z \geq -1{,}54) = 0{,}9382.$$

O resultado exato, utilizando a Binomial, será 0,9484; indicando que a solução dada pela distribuição Normal é bastante razoável. Na Figura 6.4, representamos o histograma da Binomial e a densidade da Normal utilizada na aproximação, que é baseada no Teorema Central do Limite, um importante resultado a ser apresentado no Capítulo 7. Em geral, quanto mais simétrica for a função de probabilidade da

Binomial, melhor será a aproximação. Nos casos em que certa assimetria estiver presente, valores crescentes de n fornecerão melhores resultados.

Figura 6.4: Aproximação Normal para o Modelo Binomial.

Para melhorar a aproximação, alguns autores introduzem a *correção de continuidade* no cálculo com a Normal. Esse mecanismo consiste em alterar de 0,5 unidade o valor com que se deseja calcular a probabilidade. A alteração para mais ou para menos depende, respectivamente, da probabilidade desejada excluir ou incluir a igualdade ao valor desejado. Por exemplo, teríamos,

$$P(X \geq 50) \simeq P(Y \geq 49{,}5) = P\left(Z \geq \frac{49{,}5-60}{\sqrt{42}}\right) = 0{,}9474\,;$$

$$P(X > 50) \simeq P(Y \geq 50{,}5) = P\left(Z \geq \frac{50{,}5-60}{\sqrt{42}}\right) = 0{,}9292\,.$$

Note que, com relação a Y e Z, é indiferente se a desigualdade inclui ou não o sinal de igual.

Para calcular a igualdade a um valor, digamos $X = 50$, criamos um intervalo artificial, pois com variáveis contínuas essa probabilidade seria zero.

Assim,

$$P(X = 50) \simeq P(49{,}5 \leq Y \leq 50{,}5)$$

$$\simeq P\left(\frac{50{,}5 - 60}{\sqrt{42}} \leq Z \leq \frac{49{,}5 - 60}{\sqrt{42}}\right) = 0{,}0182.$$

O cálculo exato da probabilidade fornece o valor 0,0190; mostrando, novamente, a qualidade da aproximação. □

Como ilustração adicional, apresentamos na Figura 6.5 vários histogramas obtidos a partir da simulação de uma $b(n, p)$. Em cada linha da figura o valor de p é fixado; temos assim p igual a 0,2; 0,3 e 0,5. Dentro de cada linha o valor de n vai sendo aumentado, tomando os valores 10, 30, 50 e 100.

Figura 6.5: Histogramas para valores simulados da Binomial.

Note como o histograma se aproxima de um modelo simétrico e em forma de sino (semelhante ao modelo Normal) à medida que caminhamos da esquerda para a direita (valores crescentes de n). Pode também ser notado que a convergência será mais rápida em situações em que a distribuição Binomial é aproximadamente simétrica, o que ocorre para valores de p próximos a 1/2.

Uma propriedade muito importante do modelo Normal, cuja demonstração será omitida, é aquela que garante que qualquer combinação linear de variáveis Normais independentes, também, terá distribuição Normal. Em outras palavras, se $X_1, X_2, \ldots, X_n$ formam uma seqüência de variáveis aleatórias $N(\mu_i, \sigma_i^2)$ independentes e $a_1, a_2, \ldots, a_n$, são constantes quaisquer, então $W = \sum_{i=1}^{n} a_i X_i$ terá distribuição Normal. Seus parâmetros são determinados a partir das propriedades do valor esperado e da variância, ou seja,

$$\mu_W = E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} E(a_i X_i) = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_i;$$

$$\sigma_W^2 = Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} Var(a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 .$$

Este resultado amplia, consideravelmente, o uso da Normal em várias situações, conforme pode ser notado nos exemplos a seguir.

Exemplo 6.10: Um serviço de fiscalização é criado para averiguar se garrafas de um certo refrigerante contém, de fato, o volume especificado pelo fabricante. Para tanto, 10 garrafas do produto são compradas no varejo, em várias regiões da cidade. Cada uma dessas garrafas é esvaziada e o volume de seu conteúdo, que denotaremos por V, é aferido. Uma vez obtidos os 10 valores, a média aritmética M é calculada e, se $M < 290$ mililitros (ml), a companhia é multada. Estudos na linha de produção do fabricante mostraram que variações sempre ocorrem, mesmo se as especificações forem seguidas. Por essa razão, considera-se o volume do conteúdo das garrafas como seguindo um modelo Normal, com média $\mu = 300$ ml e desvio-padrão $\sigma = 25$ ml. Gostaríamos de calcular qual é a probabilidade de que o fabricante seja multado injustamente?

A multa será injusta se, apesar de dentro das especificações, o valor de M for abaixo de 290 ml. Observe que isto pode ocorrer devido à natureza aleatória do enchimento das garrafas.

Denotando por V_i o volume da i-ésima garrada a ser aferida e supondo que o fabricante esteja dentro das especificações, temos que $V_i \sim N(300, 25^2)$, $i = 1, \ldots, 10$. A média aritmética M é dada por

$$M = \frac{V_1 + \cdots + V_{10}}{10} = \frac{1}{10}V_1 + \cdots + \frac{1}{10}V_{10},$$

que corresponde a uma combinação linear com $a_i = 1/10$, para todo i. Logo, assumindo independência entre as variáveis aleatórias V_i, $i = 1, \ldots, 10$, temos que $M \sim N(\mu_M, \sigma_M^2)$, com

$$\mu_M = \sum_{i=1}^{10} a_i \mu = \sum_{i=1}^{10} \frac{1}{10} 300 = 300;$$

$$\sigma_M^2 = \sum_{i=1}^{10} a_i^2 \sigma^2 = \sum_{i=1}^{10} \left(\frac{1}{10}\right)^2 25^2 = \frac{25^2}{10}.$$

Logo,

$$P(\text{multa}) = P(M < 290) = P\left(\frac{M - \mu_M}{\sigma_M/\sqrt{n}} < \frac{290 - 300}{25/\sqrt{10}}\right)$$

$$= P(Z < -1{,}26) = 0{,}1038.$$

Portanto, a probabilidade de que a empresa seja multada, indevidamente, será de, aproximadamente, 10 %. □

Exemplo 6.11: Uma corretora negocia títulos na Bolsa de Valores e utiliza um modelo probabilístico para avaliar seus lucros. Suas aplicações financeiras de compra e venda atingem três áreas: agricultura, indústria e comércio. Admita que o seguinte modelo representa o comportamento do lucro diário da corretora (em milhares de reais):

$$L = 2\,L_{\text{A}} + 5\,L_{\text{I}} + 3\,L_{\text{C}},$$

com $L_{\text{A}}, L_{\text{I}}$ e L_{C} representando, respectivamente, os lucros diários nos setores de agricultura, indústria e comércio. As distribuições de probabilidade dessas variáveis aleatórias são $L_{\text{A}} \sim N(3, 4)$, $L_{\text{I}} \sim N(6, 9)$ e $L_{\text{C}} \sim N(4, 16)$. Supondo independência entre os três setores, qual será a probabilidade de um lucro diário acima de 50 mil?

Conforme mencionado, a variável L, sendo uma combinação linear de Normais independentes, segue distribuição Normal com parâmetros dados por:

$$\mu_L = 2 \times 3 + 5 \times 6 + 3 \times 4 = 48;$$

$$\sigma_L^2 = 2^2 \times 4 + 5^2 \times 9 + 3^2 \times 16 = 385.$$

Então $L \sim N(48, 385)$ e, portanto,

$$P(L > 50) = P(Z > \frac{50 - 48}{\sqrt{385}}) = P(Z > 0{,}10) = 0{,}4602;$$

indicando uma alta probabilidade de lucros superiores a 50 mil. □

Exercícios da Seção 6.2:

1. Sendo $X \sim U[0, 4]$, calcule
- **a.** $P(X > 2)$.
- **b.** $P(X \geq 2)$.
- **c.** $P(1 < X < 2)$.
- **d.** $P(1 < X < 2 | X < 3)$.
- **e.** $P(X < 3 | 1 < X < 2)$.

2. Admite-se que uma pane pode ocorrer em qualquer ponto de uma rede elétrica de 10 quilômetros.
- **a.** Qual é a probabilidade da pane ocorrer nos primeiros 500 metros? E de ocorrer nos 3 quilômetros centrais da rede?
- **b.** O custo de reparo da rede depende da distância do centro de serviço ao local da pane. Considere que o centro de serviço está na origem da rede e que o custo é de R$ 200 para distâncias até 3 quilômetros, de R$ 400 entre 3 e 8 e de R$ 1.000 para as distâncias acima de 8 quilômetros. Qual é o custo médio do conserto?

3. O tempo necessário para um medicamento contra dor fazer efeito foi modelado de acordo com a densidade Uniforme no intervalo de 5 a 15 (em minutos), tendo por base experimentos conduzidos em animais. Um paciente, que esteja sofrendo dor, recebe o remédio e, supondo válido o modelo mencionado acima, pergunta-se a probabilidade da dor :
- **a.** Cessar em até 10 minutos?
- **b.** Demorar pelo menos 12 minutos?
- **c.** Durar mais de 7 minutos, sabendo-se que durou menos de 10?

4. Suponha que o valor esperado de uma variável aleatória com distribuição Uniforme contínua é 1 e a variância é igual a 1/12. Encontre a probabilidade da variável assumir valores menores que 3/4.

5. Sendo $X \sim Exp(1)$, determine:
   a. $P(0 < X < 2)$.
   b. $P(X < 2)$.
   c. $P(1 < X < 4)$.
   d. $P(X > 3)$.
   e. $P(X < 2 | X > 1)$.

6. Suponha que o tempo de vida T de um vírus exposto ao meio ambiente segue uma distribuição Exponencial com parâmetro $\lambda = 1/20$ s. Calcule a probabilidade condicional $P(T > 15 \mid T > 10)$.

7. Seja $X \sim N(4,1)$. Determine:
   a. $P(X \leq 4)$.
   b. $P(4 < X < 5)$.
   c. $P(2 \leq X < 5)$.
   d. $P(5 \leq X \leq 7)$.
   e. $P(X \leq 1)$.
   f. $P(0 \leq X \leq 2)$.

8. Para $X \sim N(90,100)$, obtenha:
   a. $P(X \leq 115)$.
   b. $P(X \geq 80)$.
   c. $P(X \leq 75)$.
   d. $P(85 \leq X \leq 110)$.
   e. $P(|X - 90| \leq 10)$.
   f. O valor de a tal que $P(90 - a \leq X \leq 90 + a) = \gamma$, $\gamma = 0{,}95$.

9. Para $X \sim N(-5, 10)$, calcule:
   a. $P(-5 < X \leq -2)$.
   b. $P(X \leq 0)$.
   c. $P(X > -6)$.
   d. $P(-7 \leq X \leq -6)$.
   e. $P(|X + 5| > 2)$.

6.3 Exercícios

1. Uma variável contínua tem densidade dada por:

$$f(x) = \begin{cases} \frac{1}{4}x, & 1 \leq x \leq 3; \\ 0, & \text{caso contrário.} \end{cases}$$

   a. Faça um gráfico da função acima e verifique que ela satisfaz as condições para ser densidade.
   b. Determine $P(X > 2)$, $P(X \geq 2)$ e $P(X = 2)$.
   c. Calcule $P(0 \leq X \leq 3/2)$, $P(X > 3/2)$ e $P(X \geq 3/2)$.
   d. Obtenha $P(3/2 < X \leq 5/2 \mid X < 2)$.

2. A densidade de probabilidade para uma variável aleatória contínua X é dada pelo gráfico a seguir:

Determine:
   a. $P(-1/4 < X < 1/4)$.
   b. $P(0 \leq X < 3/4)$.
   c. $P(-3/4 \leq X < 0)$.
   d. $P(X > 0, X > 3/4)$.
   e. $P(X > 0 \mid X < 1/2)$.
   f. $P(X \leq 3/4 \mid X > 1/2)$.
   g. Um número b tal que $P(X > b) = 1/4$.

3. Suponha que uma variável aleatória contínua tenha densidade de probabilidade dada por:

$$f(x) = \begin{cases} \frac{1}{6}x + k, & 0 < x < 3; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Qual é o valor de k?
b. Quanto vale b, tal que $P(X > b) = 5/9$?

4. O tempo adequado de troca de um amortecedor de certa marca em automóveis, sujeitos a uso contínuo e severo, pode ser considerado como uma variável contínua, medida em anos. Suponha que a função densidade é dada pela seguinte expressão:

$$f(x) = \begin{cases} \frac{1}{4}x, & 0 \leq x \leq 2; \\ \frac{1}{8}, & 2 < x \leq 6; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Verifique que a função acima é, de fato, uma densidade.
b. Qual é a probabilidade de um automóvel, sujeito às condições descritas acima, necessitar de troca de amortecedores antes de 1 ano de uso? E entre 1 e 3 anos?
c. Supondo que um automóvel está há 3 anos com o mesmo amortecedor, qual a probabilidade de que seja necessário fazer a troca antes de completar 4 anos de uso?
d. Qual é o tempo médio adequado para a troca do amortecedor desses automóveis?

5. O gráfico, a seguir, representa a densidade de uma variável aleatória contínua X.
a. Verifique que $f(x)$ representa uma densidade.
b. Escreva a expressão da função.
c. Calcule $P(X < 5/12)$.
d. Determine um número c tal que $P(X < c) = 1/2$.

6. O acréscimo anual na área atingida por uma certa praga, numa região produtora de frutas, pode ser modelado por uma variável aleatória contínua, medida em hectares (10 mil m^2), com densidade:

$$f(x) = \begin{cases} \frac{2}{3}x, & 0 < x < 1; \\ 1 - \frac{x}{3}, & 1 \leq x < 3; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Construa o gráfico dessa densidade.
b. Qual seria a probabilidade da praga atingir entre 2 e 3 hectares esse ano?
c. Que área será atingida com 50% de certeza?
d. Determine o acréscimo médio anual na área atingida pela praga.

7. Suponha que o peso de recém-nascidos (em kg) pode ser considerado uma variável aleatória com a seguinte densidade:

$$f(x) = \begin{cases} \frac{1}{10}x + \frac{1}{10}, & 0 \leq x \leq 2; \\ -\frac{3}{40}x + \frac{9}{20}, & 2 < x \leq 6; \\ 0, & \text{caso contrário.} \end{cases}$$

Qual a probabilidade de, escolhendo ao acaso uma criança, ela ter peso:
a. Inferior a 3 kg?
b. Entre 1 e 4 kg?
c. Pelo menos 3 kg?

8. Um parafuso produzido por um torno automático poderá ter uma pequena variação no seu comprimento, dada em milímetros. Afirma-se que o comportamento aleatório pode ser modelado pela seguinte função:

$$f(x) = \begin{cases} kx + \frac{15}{32}, & \frac{1}{16} \leq x \leq \frac{33}{16}; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Determine o valor de k tal que $f(x)$ seja, de fato, uma função densidade de probabilidade.

b. Para um parafuso escolhido ao acaso dentre os produzidos por esse torno, qual a probabilidade de obtermos um comprimento maior que 1 milímetro?

c. Calcule o comprimento médio desses parafusos.

9. Suponha que uma variável aletória contínua X tem densidade dada pela seguinte função:

$$f(x) = \begin{cases} \frac{1}{6}, & 0 \leq x \leq 1; \\ \frac{1}{2}x - \frac{1}{3}, & 1 < x \leq 2; \\ -\frac{1}{2}x + \frac{5}{3}, & 2 < x \leq 3; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Determine a mediana e a média.

b. Calcule a variância.

10. A função apresentada, a seguir, corresponde à densidade de uma variável aleatória contínua X.

$$f(x) = \begin{cases} \frac{1}{4}x^3, & 0 \leq x \leq 2; \\ 0, & \text{caso contrário.} \end{cases}$$

Determine:

a. $P(X > 1)$.

b. $P(X < 1/2)$.

c. $P(1/2 < X \leq 1 | X < 3/2)$.

11. Suponha que o tempo, em meses, para a recuperação de pacientes submetidos a um certo tipo de cirurgia do aparelho digestivo pode ser modelado por uma variável aleatória contínua X, cuja função densidade de probabilidade é dada por:

$$f(x) = \begin{cases} \frac{1}{3}, & 0 \leq x \leq 1; \\ -\frac{1}{12}x + \frac{5}{12}, & 1 < x \leq 5; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Determine a média e a mediana do tempo de recuperação.

b. Calcule o desvio padrão.

12. A trava de segurança de um aparelho industrial deve ser trocada com freqüência, de modo a evitar a quebra devido ao fim de sua vida útil. Estudos anteriores admitem que essa vida útil pode ser representada por uma variável aleatória contínua, assumindo valores entre 0 e 1 ano. Sua densidade é a seguinte:

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2), & 0 < x \leq 1; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Calcule a probabilidade da vida útil ser superior a 6 meses.

b. Determine a vida útil média.

13. Suponha que o comprimento de fósseis encontrados em uma certa região, dado em centímetros, pode ser representado por uma variável aleatória X com função densidade de probabilidade dada por:

$$f(x) = \begin{cases} \frac{1}{24}, & 0 \leq x \leq 12; \\ \frac{1}{192}(x - 4), & 12 < x \leq 20; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Calcule a média e a variância de X.

b. Se um museu decide comprar os fósseis encontrados pagando R$ 100,00 para os de comprimento menor que 10 centímetros e R$ 200,00 para os demais, quanto paga em média por exemplar?

14. O tempo de corrosão, em anos, de uma certa peça metálica é uma variável com densidade:

$$f(x) = \begin{cases} ax, & 0 \leq x \leq 1; \\ a, & 1 < x \leq 2; \\ -ax + 3a, & 2 < x \leq 3; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Calcule a constante a.

b. Uma peça é considerada como tendo boa resistência à corrosão se dura mais que 1,5 anos. Em um lote de 3 peças, qual a probabilidade de termos exatamente 1 delas com boa resistência?

15. O consumo de combustível de um certo automóvel é uma variável aleatória, medida em quilômetros por litro. Admita que a densidade de probabilidade dessa variável é expressa pela seguinte função:

$$f(x) = \begin{cases} x - 10, & 10 \leq x \leq 11; \\ 12 - x, & 11 < x \leq 12; \\ 0, & \text{caso contrário.} \end{cases}$$

a. Determine a média e a variância do consumo.

b. Sendo R$ 0,7 o preço do litro de combustível, qual será a média da despesa em uma viagem de 100 quilômetros com esse automóvel.

16. Sendo $X \sim U[0, 4]$, determine:

a. $P(0 < X < 2)$.

b. $P(X < 2)$.

c. $P(1 < X < 4)$.

d. $P(X > 3)$.

e. $P(X < 2 | X > 1)$.

17. Vigas de ferro são soldadas em toda sua extensão a uma estrutura metálica. Uma falha na soldagem pode aparecer com probabilidade 0,1 e, se ocorrer, será em qualquer ponto da viga com igual probabilidade. Se a viga tem especificação de comprimento igual a 6 metros, determine a probabilidade de que:

a. Sabendo-se que uma falha ocorreu, ela ser distante no máximo em 1 metro das extremidades.

b. Ocorrer falha de solda nos dois metros centrais da viga.

18. Dois amigos planejam um encontro entre 20 e 21 horas. Um deles é pontual e pretende chegar às 20:30 horas e esperar por exatos 15 minutos. O outro é mais imprevisível e poderá chegar em qualquer momento do intervalo inicialmente previsto, saindo imediatamente se não encontrar o amigo. Qual é a probabilidade deles se encontrarem? Qual é a probabilidade deles não se encontrarem por um lapso de no máximo 5 minutos?

19. Em uma empresa, o equipamento de ar condicionado trabalha continuamente, exceto quando ocorre alguma falha que causará uma interrupção e necessidade de manutenção. Vamos supor que pode haver, no máximo, uma falha por semana (7 dias), que ocorre com probabilidade 0,05. Em havendo falha, ela pode ocorrer em qualquer hora do dia (24 horas).

a. Se o expediente na empresa vai de 8 às 18 horas de segunda a sexta, qual a probabilidade de uma falha durante o expediente?

b. As falhas, durante o expediente, acarretam custos de R$ 300,00 enquanto que nos demais horários o custo é de R$ 200,00. Admita que se não houver falha, o custo é zero. Em 4 semanas, qual é o custo médio devido a falhas com o ar condicionado?

20. Seja $X \sim Exp(1/10)$, calcule:

a. $P(X \leq 5)$.

b. $P(4 < X < 6)$.

c. $P(2 \leq X < 5)$.

d. $P(X \leq 7 | X > 2)$.

e. O valor esperado de Y, sendo $Y = 3X + 2$.

f. A variância de Y.

21. Para uma Exponencial de parâmetro 1, calcule a probabilidade de sortearmos um valor que se distancie no máximo 0,5 da média. Obtenha a expressão da função de distribuição dessa variável. Qual é o valor do terceiro quartil?

22. O tempo, em minutos, de utilização de um caixa eletrônico por clientes de um certo banco, foi modelado por uma variável T com densidade Exponencial (3). Determine:

a. $P(T < 1)$.

b. $P(T > 1 | T \leq 2)$.

c. Um número a tal que $P(T \leq a) = 0{,}4$.

23. O tempo necessário para eliminar o perigo de contaminação de certo pesticida, após sua aplicação em um pomar, é uma variável aleatória Exponencial de parâmetro 2 (em anos). O maior ou menor tempo depende de fatores como chuva, vento e umidade da região. Tendo em vista esse comportamento, as autoridades sanitárias recomendam que o contato direto ou indireto com as frutas pulverizadas seja evitado por algum tempo após a aplicação. Calcule a probabilidade de uma fruta desse pomar, escolhida ao acaso, não estar mais contaminada após 1 ano da pulverização. Qual é a nossa "segurança" se aguardarmos 2 anos para consumir essas frutas?

24. Um banco faz operações via Internet e, após um estudo sobre o serviço prestado, concluiu o seguinte modelo teórico para o tempo de conexão (em minutos):

$$f(x) = \frac{1}{4}k\,e^{-\frac{1}{4}kx},\ x > 0,$$

com k sendo 1 ou 2, dependendo do cliente ser pessoa física ou jurídica. Dentre os clientes que se utilizam da Internet, a porcentagem dos que são classificados como pessoa física é estimada em 20%.

a. Sendo pessoa física, qual a probabilidade de mais de 2 minutos de conexão?
b. Sendo pessoa jurídica, qual a probabilidade de ficar conectado menos de 6 minutos?
c. Determine a probabilidade de um cliente ficar mais de 2 minutos conectado.
d. Se um cliente fica mais de 5 minutos conectado, qual a probabilidade dele ser pessoa jurídica?

25. Seja $X \sim N(5, 4)$. Determine:
a. $P(X \leq 6)$.
b. $P(7 < X < 8)$.
c. $P(2 \leq X < 5)$.
d. $P(-1 \leq X \leq 2)$.
e. $P(X \leq -1)$.
f. $P(-2 \leq X \leq -1)$.

26. Para $X \sim N(50, 81)$, obtenha:
a. $P(X \leq 75)$.
b. $P(X \geq 60)$.
c. $P(X \leq 35)$.
d. $P(85 \leq X \leq 100)$.
e. $P(|X - 40| \leq 10)$.

27. Sendo $X \sim N(5/4, 1/9)$, calcule:
a. $P(X \leq 7/5)$.
b. $P(0 < X < 6/5)$.
c. $P(X \leq 3/5)$.
d. $P(|X - 4/3| \leq 1/2)$.
e. $P(|X - 1| > 1/4)$.

28. Na distribuição $X \sim N(\mu, \sigma^2)$, encontre:
a. $P(X \geq \mu + 2\sigma)$.
b. $P(|X - \mu| \leq \sigma)$.
c. O número a tal que $P(\mu - a\sigma \leq X \leq \mu + a\sigma) = 0{,}99$.
d. O número a tal que $P(X > a) = 0{,}90$.

29. Uma clínica de emagrecimento recebe pacientes adultos com peso seguindo uma distribuição Normal de média 130 kg e desvio padrão 20 kg. Para efeito de determinar o tratamento mais adequado, os 25% pacientes de menor peso são classificados de "magros", enquanto os 25% de maior peso de "obesos". Determine os valores que delimitam cada uma dessas classificações.

30. Um teste de aptidão feito por pilotos de aeronaves em treinamento inicial requer que uma série de operações seja realizada em uma rápida sucessão. Suponha que o tempo necessário para completar o teste seja distribuído de acordo com uma Normal de média 90 minutos e desvio padrão 20 minutos.
a. Para passar no teste, o candidato deve completá-lo em menos de 80 minutos. Se 65 candidatos tomam o teste, quantos são esperados passar?
b. Se os 5% melhores candidatos são alocados para aeronaves maiores, quão rápido deve ser o candidato para que obtenha essa posição?

31. Com base em experiências anteriores, a Companhia Telefônica sabe que 10% das contas dos seus clientes em uma comunidade são pagas com atraso. Para os itens abaixo, compare a solução exata com aquela obtida através de aproximação da variável aleatória pela distribuição Normal.
a. Se 20 contas são enviadas em um dia pela Companhia Telefônica, qual é a probabilidade de que menos do que 3 sejam pagas com atraso?
b. Se 150 contas são enviadas mensalmente para a comunidade, encontre a probabilidade de que 17 ou mais sejam pagas com atraso.

32. A durabilidade de um tipo de pneu da marca *Rodabem* é descrita por uma variável aleatória Normal de média 60.000 km e desvio padrão de 8.300 km.
a. Se a *Rodabem* garante os pneus pelos primeiros 48.000 km, qual a proporção de pneus que deverão ser trocados pela garantia?
b. O que aconteceria com a proporção do item (a), se a garantia fosse para os primeiros 45.000km?
c. Qual deveria ser a garantia (em km) de tal forma a assegurar que o fabricante trocaria sob garantia no máximo 2% dos pneus?
d. Se você comprar 4 pneus *Rodabem*, qual será a probabilidade de que você utilizará a garantia (45.000 km) para trocar um ou mais destes pneus?

33. Estudos meteorológicos indicam que a precipitação pluviométrica mensal, em períodos de seca numa certa região, pode ser considerada como seguindo a distribuição Normal de média 30 mm e variância 16 mm^2.

a. Qual seria o valor da precipitação pluviométrica de modo que exista apenas 10% de probabilidade de haver uma precipitação inferior a esse valor?

b. Construa um intervalo central em torno da média que contenha 80% dos possíveis valores de precipitação pluviométrica.

c. Admitindo esse modelo correto para os próximos 50 meses, em quantos deles esperaríamos uma precipitação pluviométrica superior a 34 mm?

34. (Use o computador) Utilizando um gerador de números aleatórios, obtenha uma amostra de 100 observações de uma distribuição Binomial, com parâmetros $n = 10, 30, 50, 100$ e $p = 0{,}5$; 0,7 e 0,9. Para cada combinação de n e p, construa um histograma. Baseando-se na forma dos histogramas construídos, o que pode ser dito a respeito da aproximação para a distribuição Normal?

35. (Use o computador) Com os dados do arquivo *areas.txt* (ver Exercício 25, Capítulo 1):

a. Para cada bloco, construa um histograma para a variável Sala. Compare os gráficos. Existe diferença entre o tamanho da sala para cada bloco?

b. Observando a forma dos histogramas, que modelo teórico você sugeriria para a variável Sala, em cada bloco?

36. (Use o computador) Para o arquivo *cancer.txt* descrito no Exercício 24 do Capítulo 1, considere a variável LDH para os pacientes com pelo menos 40 anos de idade.

a. Obtenha o histograma e algumas medidas descritivas.

b. Você diria que o modelo Normal representaria bem esses dados?

37. (Use o computador) Com os dados do arquivo *aeusp.txt* (ver Exercício 26, Capítulo 1):

a. Verifique se a variável Temposp pode ser modelada por uma Normal com média 19 e desvio padrão 12.

b. Para cada região de procedência, construa um histograma para a variável Temposp. Compare os gráficos. Existe diferença entre eles? Algum palpite sobre os modelos teóricos que poderiam ser adequados?

c. Foi proposta a Uniforme contínua $[1, 10]$ para modelar a variável Resid. Qual é a sua opinião?

Capítulo 7

Inferência Estatística - Estimação

7.1 Introdução

A Inferência Estatística é um conjunto de técnicas que objetiva estudar a população através de evidências fornecidas por uma amostra. É a amostra que contém os elementos que podem ser observados e, a partir daí, quantidades de interesse podem ser medidas.

Para ilustrar as idéias que discutiremos nesse capítulo, considere que é de interesse estudar a proporção de alunos, em uma escola do ensino médio, que pretendem fazer vestibular. Para tanto, selecionamos uma amostra de alunos e perguntamos a eles sobre suas intenções futuras de estudo. Com o intuito de obter alguma indicação do valor da proporção na escola como um todo, podemos usar a proporção dos que pretendem prosseguir os estudos no grupo selecionado. Suponha que a escola tenha 1000 alunos e escolhemos 20 para a amostra. Essa escolha pode ser em uma mesma classe ou espalhada entre os alunos das três séries ou, ainda, realizada de tal modo a garantir igual presença de meninos e meninas no grupo, independentemente da série cursada. Uma forma simples de escolher é associar um número a cada um dos 1000 alunos, colocar todos esses números numa lista e sortear 20 números. Os alunos correspondentes aos números sorteados formariam a amostra. Suponha que você realize o sorteio dessa forma e um amigo seu, desconhecendo sua iniciativa, repita o mesmo procedimento. Você acha que as amostras sorteadas por você e por seu amigo serão as mesmas? Parece intuitivo assumir que não. Queremos enfatizar que, se realizarmos várias vezes a amostragem descrita, provavelmente obteremos amostras compostas por alunos diferentes. Uma questão que surge agora é: apesar de diferentes, podemos ter respostas próximas ou iguais nas diversas amostras? A resposta é afirmativa e estará subjacente às idéias que desenvolveremos neste capítulo.

Resumindo a discussão do parágrafo anterior, podemos dizer que devido à natureza aleatória, geralmente envolvida no procedimento amostral, não podemos garantir que repetições de amostras produzam sempre resultados idênticos. Assim, ao coletarmos uma amostra, não podemos prever antecipadamente seu resultado.

Em outras palavras, todas as quantidades associadas à amostra terão caráter aleatório e, portanto, devem receber tratamento probabilístico.

Um ponto importante a destacar é que, se a população inteira entrar na amostra, temos, na prática, toda a informação possível e não há surpresas na amostragem, ou seja, não há aleatoriedade envolvida. Por exemplo, se os 1000 alunos da escola mencionada acima forem entrevistados, teremos o valor exato da proporção dos que desejam continuar os estudos na universidade. Nesse caso, toda a população faz parte da amostra e o resultado obtido irá ser sempre o mesmo, não importando quantas vezes repetimos a coleta da amostra. É claro que estamos supondo que os alunos não trocam de opinião entre as coletas e, portanto, como todos os alunos sempre entram na amostra, a proporção obtida se mantém.

Neste capítulo, formalizaremos alguns conceitos relacionados a um ramo da Inferência Estatística denominado *estimação*. Estudaremos combinações dos valores de amostras aleatórias, objetivando a obtenção de informações a respeito de características de interesse na população.

Para eliminar ambigüidades e confusões de notação, vamos representar uma amostra de tamanho n, a ser retirada da população, por $(X_1, X_2, ..., X_n)$.

Exemplo 7.1: Uma empresa fabrica 100 equipamentos eletrônicos por semana e deseja verificar como se comporta a resistência desse tipo de equipamento em relação à alteração de voltagem. Um teste, planejado pelo controle de qualidade da empresa, consiste em produzir sucessivas alterações padronizadas de voltagem e observar o efeito no aparelho. Serão considerados com boa resistência os aparelhos que passarem no teste. Como esses testes são demorados e demandam custos expressivos, apenas 5 desses aparelhos serão testados. Que cuidados precisamos ter na escolha e na interpretação dos resultados?

Os 5 aparelhos escolhidos precisam ser "representativos" da produção, ou seja, a amostra precisa representar bem a população de aparelhos produzidos. Assim, se questões referentes a operadores, máquinas utilizadas ou, até mesmo, o dia em que foram produzidos tiverem efeito na qualidade do aparelho, elas precisam ser consideradas na amostra. Uma alternativa seria o sorteio, por exemplo, de um aparelho a cada dia, tentando não repetir operadores ou máquinas utilizadas. Além disso, fazer alguma proposta de mudança no processo de produção, baseando-se apenas nos resultados de uma semana, parece ser precipitado. O mais indicado seria coletar amostras em várias semanas.

Uma vez escolhido o esquema de amostragem, a cada elemento da amostra podem ser atribuídos valores 0 ou 1 dependendo, respectivamente, do aparelho ser classificado como tendo má ou boa resistência às alterações de

voltagem. A amostra $(X_1, X_2, ..., X_5)$ poderia ter a resposta (0, 1, 1, 1, 1) em uma coleta e numa outra (1, 1, 0, 1, 0). □

Exemplo 7.2: Duvida-se da "honestidade" de um dado e decide-se lançá-lo 10 vezes antes de utilizá-lo em um jogo. Os resultados obtidos foram: 1, 5, 1, 4, 1, 2, 3, 3, 2 e 3. A que conclusão chegamos?

A freqüência de ocorrências de cada face é apresentada na tabela abaixo:

Face	1	2	3	4	5	6
Freqüência	3	2	3	1	1	0
Freq. relativa	0,3	0,2	0,3	0,1	0,1	0

Em sendo o dado equilibrado, as freqüências de ocorrência de todas as faces deveriam ser próximas. Entretanto, a amostra coletada parece indicar um certo desbalanceamento do dado, favorecendo valores pequenos. É bom notar que mesmo um dado honesto pode produzir a amostra acima. Talvez uma amostra maior pudesse corrigir o desvio encontrado, mas, baseado no que foi informado, o melhor seria não jogar com esse dado! □

Exemplo 7.3: No primeiro dia, após mudar para um novo bairro, você decide perguntar às pessoas, no ponto de ônibus, quanto tempo se espera para o ônibus chegar. As 15 pessoas presentes forneceram os seguintes números (em minutos): 5, 10, 5, 15, 12, 12, 10, 15, 20, 15, 20, 12, 8, 10 e 10. Uma demora de 10 minutos parece ser inevitável?

É claro que as pessoas deram suas opiniões baseadas em experiências anteriores, que devem ser diferentes entre si. É possível, também, que algumas delas sejam mais atentas que outras na questão da demora, além do que não devem chegar todas no mesmo horário ao ponto. Algumas, talvez, levem em conta o comportamento das últimas semanas para dar a opinião, outras apenas o dia anterior. Também, quem sabe as pessoas pessimistas peguem o dia de maior demora e as otimistas o de menor. Dessa maneira, a subjetividade da resposta é fruto da informalidade e imprecisão da pergunta. Das informações obtidas, temos média igual a 11,6; moda 10 e mediana igual a 12. Assim, num primeiro momento, parece ser razoável acreditar em espera pouco acima de 10 minutos. □

Exercícios da Seção 7.1:

1. Liste as idades de cinco dos seus amigos. Escreva cada um desses números em um pequeno pedaço de papel, de igual tamanho, e coloque-os dentro de um envelope. Antes de cada retirada, chacoalhe vigorosamente o envelope e feche os olhos.

a. Qual é a intenção de toda a "engenharia" descrita acima?

b. Repita três vezes o seguinte procedimento: retire de uma vez três papéis do envelope e anote seus números. Comente sobre as três trincas de números encontradas.

c. Repita três vezes o seguinte procedimento: retire um dos papéis do envelope, anote o número e devolva-o ao envelope. Faça mais duas retiradas nos mesmos moldes. Comente sobre as três trincas encontradas.

d. Que diferenças existem nos procedimentos descritos em (b) e (c)?

2. Deseja-se sortear 100 crianças entre 4 e 10 anos, num certo bairro, para uma pesquisa sobre saúde bucal. Foram propostas três alternativas para a coleta:

I: Um sorteio aleatório, realizado entre as crianças listadas pelo posto de saúde do bairro.

II: Um sorteio aleatório de casas do bairro e, em seguida, uma escolha aleatória de uma criança de cada casa sorteada, se houver.

III: Escolhe-se, ao acaso, um dia de semana em uma das escolas do bairro. Nessa ocasião, 100 crianças são sorteadas dentre as várias classes, com alunos na faixa etária de interesse.

Comente as diferenças e dificuldades de cada alternativa.

7.2 Parâmetros, Estimadores e Estimativas

Para formalizar as idéias que serão apresentadas neste capítulo, precisamos definir alguns conceitos.

Definição 7.1: Parâmetro

As quantidades da população, em geral desconhecidas, sobre as quais temos interesse, são denominadas *parâmetros* e, usualmente, representadas por letras gregas tais como θ, μ e σ, entre outras. □

Definição 7.2: Estimador e estimativa

À combinação dos elementos da amostra, construída com a finalidade de representar, ou estimar, um parâmetro de interesse na população, denominamos *estimador*. Em geral, denotamos os estimadores por símbolos com o acento circunflexo: $\widehat{\theta}$, $\widehat{\mu}$, $\widehat{\sigma}$, etc. Aos valores numéricos assumidos pelos estimadores denominamos *estimativas pontuais* ou simplesmente *estimativas*. □

A notação utilizada para a média de uma população é μ, acrescida de um subscrito, se houver possibilidade de confusão sobre a que população ou variável

nos referimos. Por exemplo, μ_X e μ_A são usados para indicar, respectivamente, média de X referente a uma certa população e média da população A referente a uma certa variável. Também é usual considerar σ para indicar o desvio padrão da população, aplicando-se os mesmos comentários anteriores sobre a utilização de subscritos. Outros parâmetros de interesse já não têm uma uniformidade de notação entre os diversos autores e tentaremos utilizar aquela que nos pareça mais intuitiva.

Notamos que um estimador, digamos $\widehat{\theta}$, é uma função das variáveis aleatórias constituintes da amostra, isto é, $\widehat{\theta} = f(X_1, X_2, ..., X_n)$. Logo, um estimador também é uma variável aleatória. A correspondente distribuição de probabilidade formará a base das argumentações probabilísticas utilizadas na extrapolação da informação da amostra para os parâmetros da população.

Exemplo 7.4: Estamos interessados na média das alturas de jovens com idade entre 15 e 18 anos, nascidos na região sudeste do país. Vamos coletar uma amostra e usá-la para tirar conclusões. Para facilitar a discussão, suponha que a amostra tenha somente 10 jovens, escolhidos ao acaso dentre a população mencionada.

A população é composta por todos os jovens com idade entre 15 e 18 anos, nascidos na região sudeste. O parâmetro de interesse é a altura média desses jovens, representada por μ. A amostra $(X_1, X_2, ..., X_{10})$ será obtida e, com base nela, vamos dizer algo a respeito de μ. O que precisamos resolver é que função dos valores amostrais utilizaremos para essa tarefa, isto é, qual será o estimador. Apresentamos a seguir algumas opções:

$$\widehat{\mu}_1 = f_1(X_1, ..., X_{10}) = \frac{(\text{mínimo} + \text{máximo})}{2};$$
$$\widehat{\mu}_2 = f_2(X_1, ..., X_{10}) = X_1;$$
$$\widehat{\mu}_3 = f_3(X_1, ..., X_{10}) = \frac{X_1 + \cdots + X_{10}}{10}.$$

Poderíamos listar outros estimadores, mas os três apresentados são suficientes para ilustrar nossa discussão. Inicialmente, vamos esclarecer o significado de cada um deles. O estimador $\widehat{\mu}_1$ é a média aritmética entre os valores mínimo e máximo da amostra e $\widehat{\mu}_2$ é, simplesmente, o primeiro valor sorteado na amostra. Finalmente, $\widehat{\mu}_3$ é a média dos valores da amostra, ou seja, a média amostral. Apresentamos, a seguir, os valores observados na amostra e as respectivas estimativas obtidas com os estimadores definidos acima.

Amostra (em metros): 1,65; 1,57; 1,72; 1,66; 1,71; 1,74; 1,81; 1,68; 1,60 e 1,77.

Estimativas:

$$\widehat{\mu}_{1_{obs}} = \frac{(1{,}57 + 1{,}81)}{2} = 1{,}69\,;$$
$$\widehat{\mu}_{2_{obs}} = 1{,}65\,;$$
$$\widehat{\mu}_{3_{obs}} = \frac{1{,}65 + 1{,}57 + \cdots + 1{,}77}{10} = \frac{16{,}91}{10} = 1{,}69\,.$$

Apesar desses números, calculados para uma amostra particular, não serem muito distintos uns dos outros, não devemos escolher o estimador olhando, apenas, se a estimativa correspondente é razoável. Como decidir qual deles usar? Veremos, adiante, que esta questão é resolvida, estudando-se as propriedades dos diversos estimadores. É sempre bom lembrar que nenhum de nós sabe o verdadeiro valor da média na população, pois se ele fosse conhecido, que sentido teria coletar a amostra? □

Exemplo 7.5: Para detectar o apoio popular a um projeto governamental de reforma agrária, foram entrevistadas 400 pessoas espalhadas em várias capitais. A amostra contém as 400 respostas que consistem de *sim* (para aqueles que concordam com o projeto) e *não* (para os que discordam).

Para formalizar o problema, inicialmente caracterizamos a população de interesse como aquela formada pelos habitantes adultos do país. A informação desejada é a proporção das pessoas que concordam com o referido projeto, isto é, o parâmetro de interesse é p: *proporção dos que concordam com o projeto.*

A amostra pode ser pensada como o vetor de variáveis aleatórias $(X_1, X_2, ..., X_{400})$, cada uma delas seguindo um modelo Bernoulli, ou seja, assumindo valor 1 para *sucesso* (resposta *sim*) e 0 para *fracasso* (resposta *não*).

É intuitivo considerar como estimador a proporção amostral dos que concordam. Assim tomamos

$$\widehat{p} = \frac{\textit{número dos entrevistados que aprovam o projeto}}{400}$$

que, tendo em vista as variáveis de Bernoulli, pode ser escrito como:

$$\widehat{p} = \frac{X_1 + X_2 + \cdots + X_{400}}{400}\,.$$

Como veremos adiante, esse estimador além de intuitivo tem boas propriedades. □

Suponha, como antes, que uma amostra de tamanho n é retirada da população e representada pelo conjunto de variáveis aleatórias $(X_1, X_2, ..., X_n)$.

Denote os parâmetros média, variância e proporção de certa característica na população por μ, σ^2 e p, respectivamente. Os estimadores "naturais" para estas quantidades são as correspondentes média, variância e proporção calculadas na amostra. Representando-os, respectivamente, por $\overline{X}$, $\widehat{\sigma}^2$ e $\widehat{p}$, temos

média

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}\,;$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\,;$$

$$\widehat{p} = \frac{\textit{número de itens com a característica na amostra}}{n}\,.$$

proporção

Note que cada um dos estimadores apresentados depende dos valores pertencentes à amostra aleatória $(X_1, \ldots, X_n)$. Como veremos no decorrer desta seção, os estimadores $\overline{X}$ e $\widehat{p}$, além de serem intuitivos, têm as boas propriedades que serão definidas adiante. No entanto, com respeito à $\widehat{\sigma}^2$, uma alteração na sua expressão será necessária para que satisfaça uma dessas importantes propriedades.

Exemplo 7.6: Para estudar o nível de colesterol em uma população de esportistas, coletamos uma amostra de 10 jovens atletas, obtendo os seguintes valores: 180, 196, 185, 165, 190, 195, 180, 176, 165 e 195.

Vamos definir nosso interesse como sendo o nível médio de colesterol e, assumindo que não temos acesso à toda a população, *estimaríamos* o parâmetro μ (valor desconhecido da população) pela média amostral calculada com os valores dados, isto é,

$$\overline{x}_{obs} = \frac{180 + 196 + 185 + \cdots + 176 + 165 + 195}{10} = 182{,}7.$$

Portanto, a amostra, através do estimador $\overline{X}$, fornece para o parâmetro μ a estimativa 182,7. O limite de colesterol para pessoas sadias é 200, isto é, acima desse valor o indivíduo aumenta o seu risco de ter uma complicação cardíaca. A amostra forneceu um valor relativamente baixo, indicando que as pessoas que praticam esportes, aparentemente, estão mais protegidas de complicações do coração.

Tendo em vista que a população em estudo é constituída de jovens atletas, um nível de colesterol acima de 190 poderia ser considerado preocupante e indicativo para um acompanhamento médico mais freqüente. Dessa forma, suponha que classifiquemos como tendo *taxa alta* os atletas com valores acima de

190 e *taxa baixa*, os demais. Sendo X_i o nível de colesterol do i-ésimo atleta escolhido, definimos

$$Y_i = \begin{cases} 1, & \text{se } X_i > 190; \\ 0, & \text{se } X_i \leq 190. \end{cases}$$

Assim, Y_i será 1 para as taxas altas e 0 para as baixas. As quantidades $Y_1, \ldots, Y_{10}$ também são variáveis aleatórias, uma vez que elas assumem o valor 0 ou 1 dependendo do valor assumido por X_i, que, por sua vez, é uma variável aleatória. Para os dados apresentados, podemos construir a tabela:

i	1	2	3	4	5	6	7	8	9	10
X_i	180	196	185	165	190	195	180	176	165	195
Y_i	0	1	0	0	0	1	0	0	0	1

A proporção p de atletas com taxa de colesterol alta será estimada pela proporção de taxas altas encontradas na amostra, $\widehat{p}$. A estimativa obtida é:

$$\widehat{p}_{obs} = \frac{Y_1 + Y_2 + \cdots + Y_{10}}{10} = \frac{0 + 1 + \cdots + 1}{10} = 0{,}3.$$

Portanto, baseado na amostra disponível, assumiremos que 30% de todos os atletas têm taxa relativamente alta de colesterol, indicando a necessidade de acompanhamento médico. □

Exemplo 7.7: Foi coletada uma amostra de pacientes, sofrendo de um certo tipo de câncer, para se ter uma idéia da variabilidade da área atingida pela doença. Para 12 pacientes sorteados ao acaso mediu-se, através de aparelhos, o tamanho dos tumores observados. Os dados foram os seguintes (em cm^2): 3,52; 4,45; 3,85; 4,32; 6,12; 5,88; 4,08; 5,91; 4,50; 4,86; 5,48 e 5,10.

Tendo em vista que se deseja estudar a variabilidade, vamos considerar como parâmetro de interesse a variância σ^2. Para estimador, considere duas opções:

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2;$$

$$\widehat{\sigma}_2^2 = \left(\frac{\text{máximo} - \text{mínimo}}{2}\right)^2.$$

A primeira escolha é a variância do conjunto de dados que foi observado, enquanto que o segundo estimador proposto é a semi-amplitude de valores obtidos na amostra. Vamos calcular suas estimativas:

$$\widehat{\sigma}_{1_{obs}}^2 = \frac{1}{12}[(3{,}52 - 4{,}84)^2 + \cdots + (5{,}10 - 4{,}84)^2] = 0{,}67;$$

$$\widehat{\sigma}_{2_{obs}}^2 = \left(\frac{\text{máximo} - \text{mínimo}}{2}\right)^2 = \left(\frac{6{,}12 - 3{,}52}{2}\right)^2 = \left(\frac{2{,}60}{2}\right)^2 = 1{,}69.$$

Esses números, apesar de bem distintos, dão idéia da dispersão de valores que podem ser encontrados no tamanho dos tumores. Veremos mais adiante que um estimador "melhor" e, portanto, uma estimativa melhor pode ser obtida. □

Como vimos nos exemplos acima, mais de uma função da amostra pode ser proposta para estimar o parâmetro de interesse. Para facilitar a escolha entre tais estimadores, torna-se importante verificar se possuem algumas das propriedades que serão definidas a seguir.

Definição 7.3: Vício

Um estimador $\widehat{\theta}$ é *não viciado* ou *não viesado* para um parâmetro θ se $E(\widehat{\theta}) = \theta$. Em outras palavras, um estimador é não viciado se o seu valor esperado coincide com o parâmetro de interesse. □

Definição 7.4: Consistência

Um estimador $\widehat{\theta}$ é *consistente*, se, à medida que o tamanho da amostra aumenta, seu valor esperado converge para o parâmetro de interesse e sua variância converge para zero. Ou seja, $\widehat{\theta}$ é consistente se as duas propriedades seguintes são satisfeitas:

estimador

i) $\lim_{n\to\infty} E(\widehat{\theta}) = \theta$; parâmetro

ii) $\lim_{n\to\infty} \text{Var}(\widehat{\theta}) = 0.$

□

Note que, na definição de consistência, estamos implicitamente usando o fato que o estimador depende de n, o tamanho da amostra. Na definição do vício, o resultado deve valer para qualquer que seja n, isto é, $E(\widehat{\theta}) = \theta$, para todo n. Na definição da consistência, o estimador necessita ser não viciado apenas para valores grandes de n.

Exemplo 7.8: Considere que, numa certa população, uma variável aleatória X assuma os valores 0, 10, 20 e 30 com porcentagens 20%, 30%, 30% e 20%, respectivamente. Através da função de probabilidade, podemos calcular a média e a variância da população, nesse caso $\mu = 15$ e $\sigma^2 = 105$. Entretanto, para efeito

didático, imagine que μ não fosse conhecido e desejássemos obter informações a seu respeito através de uma amostra de tamanho 2, com reposição. Vamos identificar as possíveis amostras selecionadas.

A amostra sorteada será o par (X_1, X_2), com X_1 e X_2 sendo variáveis aleatórias independentes com a mesma distribuição da variável X. Assim,

X_1	0	10	20	30
p_i	0,2	0,3	0,3	0,2

e

X_2	0	10	20	30
p_i	0,2	0,3	0,3	0,2

.

As possíveis amostras são as seguintes: (0, 0), (0, 10), (0, 20), ... e (30, 30). São ao todo 16 pares, mas não são equiprováveis! Algumas amostras têm maior probabilidade de ocorrência do que outras. Por exemplo, a amostra (0, 10) tem probabilidade 0,06 enquanto que (10, 20) tem 0,09.

Para estimar o valor da média μ na população, considere os seguintes estimadores:

$$\widehat{\mu}_1 = f_1(X_1, X_2) = X_1\,;$$

$$\widehat{\mu}_2 = f_2(X_1, X_2) = \overline{X} = \frac{X_1 + X_2}{2}\,.$$

Como se comportam esses estimadores?

A função de probabilidade de X_1 já foi apresentada e podemos calcular sem dificuldades seu valor esperado. Como X_1 tem a mesma distribuição de X, seus valores esperados são iguais e, portanto, obtemos $E(X_1) = 15$.

Para a variável aleatória $\overline{X}$, não é difícil verificar que sua função de probabilidade é:

$\overline{X}$	0	5	10	15	20	25	30
p_i	0,04	0,12	0,21	0,26	0,21	0,12	0,04

.

Calculando o valor esperado de $\overline{X}$, obtemos

$$E(\overline{X}) = 0 \times 0{,}04 + 5 \times 0{,}12 + \cdots + 30 \times 0{,}04 = 15\,.$$

Concluímos que os estimadores $\widehat{\mu}_1$ e $\widehat{\mu}_2$, conforme definidos, são não viciados para estimar a média populacional de X. Seriam também consistentes?

O estimador $\widehat{\mu}_1$, que é igual à X_1, tem variância igual à variância de X, no caso 105, e não se altera com o tamanho da amostra. Em outras palavras, a variância será sempre 105, não importando quão grande for a amostra. Logo $\widehat{\mu}_1$ não será um estimador consistente para μ. Quanto à $\widehat{\mu}_2$, por ser a média amostral, é o quociente da soma de valores amostrados pelo tamanho da amostra e, dessa

forma, será influenciado pelo tamanho da amostra. Conforme veremos adiante, a média amostral $\overline{X}$ será um estimador consistente para a média populacional. □

Exemplo 7.9: Suponha que é sabido que uma certa característica X, na população, tem média μ e variância σ^2. Uma amostra aleatória de tamanho n, representada por $(X_1, X_2, ..., X_n)$ é obtida para estimar o parâmetro μ.

Considere o estimador $\widehat{\mu}_1 = \overline{X}$. Assumimos que o vetor amostral $(X_1, X_2, ..., X_n)$ é constituído de variáveis aleatórias independentes e todas com a mesma distribuição da variável X, isto é, $X_1, X_2, \ldots, X_n$ seguem algum modelo (que não foi especificado) com média μ e variância σ^2. Com o auxílio das propriedades da esperança e da variância temos

$$E(\widehat{\mu}_1) = E(\overline{X}) = E(\frac{X_1 + X_2 + \cdots + X_n}{n}) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n\,\mu}{n} = \mu\,.$$

Para a variância temos,

$$Var(\widehat{\mu}_1) = Var(\frac{X_1 + X_2 + \cdots + X_n}{n}) = \frac{\sum_{i=1}^{n} Var(X_i)}{n^2} = \frac{n\,\sigma^2}{n^2} = \frac{\sigma^2}{n}\,.$$

Portanto, mostramos que a média amostral é um estimador não viciado para a média populacional μ e, como $\text{Var}\,(\widehat{\mu}_1) = \sigma^2/n$ tende a 0 conforme n cresce, concluímos também que $\overline{X}$ é um estimador consistente para μ.

Considere agora que especificamos o modelo de X como sendo Normal, isto é, temos $X \sim N(\mu,\ \sigma^2)$. Os resultados apresentados acima para $\overline{X}$ permanecem válidos, pois foram desenvolvidos sem nenhuma particularidade de modelo. Suponha que um outro estimador é proposto: $\widehat{\mu}_2 = \text{mediana}(X_1, ..., X_n)$. A justificativa para o uso de $\widehat{\mu}_2$ deve-se ao fato de que o modelo Normal é simétrico e a mediana, assim como a média, é uma medida de tendência central.

Os cálculos referentes à $\widehat{\mu}_2$ são um pouco mais elaborados e não serão desenvolvidos aqui. Entretanto, pode-se demonstrar que $E(\widehat{\mu}_2) = \mu$ e $Var(\widehat{\mu}_2) = (\pi/2)(\sigma^2/n)$, de tal forma que esse estimador é também não viciado e consistente para μ . □

Exemplo 7.10: Supondo uma amostra $(X_1, \ldots, X_n)$ obtida de uma população com média μ e variância σ^2, um estimador "natural" da variância foi apresentado anteriormente e motivado pelo cálculo da variância de um conjunto de dados. Esse estimador foi denotado por $\widehat{\sigma}_1^2$ e é viciado para σ^2, uma vez que:

$$
\begin{aligned}
E(\widehat{\sigma}_1^2) &= \frac{1}{n}E\{\sum_{i=1}^{n}(X_i - \overline{X})^2\} \\
&= \frac{1}{n}E\{\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2\} \\
&= \frac{1}{n}E\{\sum_{i=1}^{n}(X_i - \mu)^2\} - \frac{1}{n}E\{n(\overline{X} - \mu)^2\} \\
&= \frac{1}{n}\sum_{i=1}^{n}E(X_i - \mu)^2 - E(\overline{X} - \mu)^2 \\
&= \frac{1}{n}n\sigma^2 - \frac{1}{n}\sigma^2 \\
&= (\frac{n-1}{n})\sigma^2.
\end{aligned}
$$

Como é imediato verificar, o quociente $(n-1)/n$ nunca será 1, exceto no limite quando n tende a infinito. Podemos eliminar o vício multiplicando $\widehat{\sigma}_1^2$ por n e dividindo por $(n-1)$. Assim, definimos um novo estimador

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

que é não viciado para σ^2. Para seu cálculo, podemos usar a expressão alternativa

$$S^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\Big).$$

Esse estimador recebe o nome de *variância amostral* e será sempre denotado por S^2 para distinguir de outros estimadores denotados genericamente por $\widehat{\sigma}^2$. □

Note que a variância ou o desvio padrão de um estimador fornece uma idéia da sua precisão. Por isso, é comum denominar o desvio padrão de um estimador de *erro padrão*. Quando dois estimadores forem consistentes e não viciados para um mesmo parâmetro, o erro padrão é considerado para decidir qual é o mais preciso. Neste contexto, o conceito de *eficiência* é apresentado a seguir.

Definição 7.5: Eficiência

Dados dois estimadores $\widehat{\theta}_1$ e $\widehat{\theta}_2$, não viciados para um parâmetro θ, dizemos que $\widehat{\theta}_1$ é mais eficiente do que $\widehat{\theta}_2$ se $Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2)$. □

Exemplo 7.11: No Exemplo 7.9, no caso de distribuição Normal, verificamos que os estimadores $\widehat{\mu}_1 = \overline{X}$ e $\widehat{\mu}_2 = \text{mediana}(X_1, ..., X_n)$ são não viciados e suas variâncias foram calculadas. Então,

$$\frac{Var(\widehat{\mu}_1)}{Var(\widehat{\mu}_2)} = \frac{\sigma^2/n}{(\pi/2)\sigma^2/n} = \frac{2}{\pi} = 0{,}63 < 1 \Rightarrow Var(\widehat{\mu}_1) < Var(\widehat{\mu}_2),$$

e concluímos que $\widehat{\mu}_1$ é mais eficiente do que $\widehat{\mu}_2$. □

Na tabela a seguir, apresentamos estimadores de μ, p e σ^2.

Tabela 7.1: Estimadores para média, proporção e variância.
(*) a consistência não foi demonstrada no texto mas é válida

Parâmetro	Estimador	Propriedades
μ	$\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$	Não viciado e consistente
p	$\widehat{p} = \frac{\text{freqüência amostral com a característica}}{n}$	Não viciado e consistente
σ^2	$S^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\Big)$	Não viciado e consistente(*)
σ^2	$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$	Viciado e consistente(*)

Exercícios da Seção 7.2:

1. Foram sorteadas 15 famílias com filhos num certo bairro e observado o número de crianças de cada família, matriculadas na escola. Os dados foram: 1, 1, 2, 0, 2, 0, 2, 3, 4, 1, 1, 2, 0, 0, e 2. Obtenha as estimativas correspondentes aos seguintes estimadores da média de crianças na escola nesse bairro:

$$\widehat{\mu}_1 = \frac{(\text{mínimo} + \text{máximo})}{2};$$
$$\widehat{\mu}_2 = \frac{(X_1 + X_2)}{2};$$
$$\widehat{\mu}_3 = \overline{X}.$$

Qual deles é o melhor estimador da média e por quê?

2. Para se estudar a variabilidade em um teste de Inglês (notas de 0 a 5), foram sorteados 16 alunos de uma escola e suas notas anotadas: 0, 1, 2, 1, 2, 3, 3, 2, 3, 3, 4, 5, 1, 3, 2 e 3. Para estimar a variância foram propostos os estimadores:

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\,;$$

$$\widehat{\sigma}_2^2 = \frac{(\text{máximo})^2 - (\text{mínimo})^2}{2}\,;$$

$$\widehat{\sigma}_3^2 = S^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\Big).$$

Obtenha as estimativas e discuta qual é melhor.

3. O número de reclamações que chegam por hora à uma Central de Atendimento do Consumidor foi anotado para uma amostra de algumas horas escolhidas ao acaso. Deseja-se organizar o serviço de modo a atender, imediatamente, 90% das chamadas que chegam. Determine uma estimativa para o número de funcionários necessários, se a amostra observada foi a seguinte: 2, 2, 3, 1, 4, 5, 3, 4, 5, 6, 5, 3, 4, 3, 4, 2, 4, 3, 5, 2, 1, 6, 2, 3 e 4.

4. Um ônibus passa por um determinado ponto em intervalos regulares (em minutos inteiros) que você, por ser novo no bairro, desconhece. Ao chegar a esse ponto você encontra duas pessoas e resolve perguntar a elas sobre seu ônibus. Uma delas diz que está no ponto há quase 10 minutos e o ônibus não passou. A outra está há cerca de 40 minutos e já viu passar dois desses ônibus. Faça uma estimativa da demora para passar o seu ônibus.

5. Um fabricante deseja estudar a duração de baterias que são utilizadas em relógios de pulso. Uma amostra de vários lotes fabricados por uma mesma companhia foi submetida a testes acelerados e produziram os seguintes tempos de duração (em anos): 1,2; 1,4; 1,7; 1,3; 1,2; 2,3; 2,0; 1,5; 1,8; 1,4; 1,6; 1,5; 1,7; 1,5 e 1,3. Determine estimativas para a média e a variância do tempo de duração dessa pilhas. Para a variância, use os estimadores da Tabela 7.1.

7.3 Distribuições Amostrais

Vimos que estimadores são funções de variáveis aleatórias e, portanto, eles também são variáveis aleatórias. Nesta seção vamos estudar a distribuição de probabilidade de alguns dos estimadores mais utilizados. Iniciamos com um exemplo simples, em que não é difícil calcular explicitamente a função de probabilidade dos estimadores de interesse.

Exemplo 7.12: Um jogo consiste em lançar uma moeda honesta 3 vezes. Para cada lançamento, se sair cara você ganha 1 ponto, caso saia coroa, você perde um

ponto. Podemos modelar essa situação através de uma variável X que, em uma população, pode assumir os valores -1 e 1, com probabilidades iguais. Para uma amostra aleatória e independente de tamanho 3, vamos determinar as funções de probabilidade dos estimadores $\overline{X}$ e S^2.

Um cálculo simples fornece o valor da média e variância de X, obtendo-se, respectivamente, 0 e 1. O vetor amostral (X_1, X_2, X_3) é constituído de variáveis aleatórias independentes e identicamente distribuídas com função de probabilidade igual à de X. A tabela, a seguir, apresenta as possíveis amostras, respectivas probabilidades e valores de $\overline{X}$ e S^2.

(X_1, X_2, X_3)	probabilidade	$\overline{X}$	S^2
$(-1,-1,-1)$	1/8	-1	0
$(-1,-1,\ 1)$	1/8	$-1/3$	4/3
$(-1,\ 1,-1)$	1/8	$-1/3$	4/3
$(-1,\ 1,\ 1)$	1/8	1/3	4/3
$(\ 1,-1,-1)$	1/8	$-1/3$	4/3
$(\ 1,-1,\ 1)$	1/8	1/3	4/3
$(\ 1,\ 1,-1)$	1/8	1/3	4/3
$(\ 1,\ 1,\ 1)$	1/8	1	0

Os valores da tabela foram obtidos através dos cálculos usuais. Por exemplo, para a amostra $(-1, 1, -1)$, temos

$$\overline{x}_{obs} = \frac{-1+1-1}{3} = -1/3\,;$$

$$s^2_{obs} = \frac{(-1)^2 + (1)^2 + (-1)^2 - 3(-1/3)^2}{(3-1)} = 4/3.$$

Baseando-se na tabela anterior, podemos construir as distribuições dos estimadores, dadas por:

$$\begin{array}{c|cccc} \overline{X} & -1 & -1/3 & 1/3 & 1 \\ \hline p_i & 1/8 & 3/8 & 3/8 & 1/8 \end{array} \quad \text{e} \quad \begin{array}{c|cc} S^2 & 0 & 4/3 \\ \hline p_i & 1/4 & 3/4 \end{array}.$$

Os valores esperados podem ser calculados facilmente:

$$E(\overline{X}) = (-1)\times 1/8 + (-1/3)\times 1/8 + 1/3\times 1/8 + 1\times 1/8 = 0\,;$$

$$E(S^2) = 0\times 1/4 + 4/3\times 3/4 = 1.$$

Dessa forma, uma vez que $E(\overline{X}) = 0 = E(X)$ e $E(S^2) = 1 = Var(X)$, ambos os estimadores são não viciados para os respectivos parâmetros estimados. □

No exemplo anterior, pudemos enumerar as possíveis amostras e, assim, obter a função de probabilidade dos estimadores de interesse. Nem sempre isso será possível. Por exemplo, se X tivesse distribuição Uniforme entre -1 e 1, X_1 e X_2 continuariam sendo independentes e identicamente distribuídas com função densidade $U[-1, 1]$. Entretanto, dada a natureza contínua das variáveis aleatórias envolvidas, não seria viável enumerar todas as amostras possíveis, o que impede a obtenção das densidades para a média e para a variância amostral pelo método apresentado no Exemplo 7.12. Não vamos entrar em detalhes aqui, mas o que é importante ressaltar, ao nível deste texto, é que obter a distribuição de probabilidade dos estimadores é um problema relevante, importante e, não raras vezes, difícil. Tendo em vista essas questões, vamos nos concentrar em discutir a distribuição de $\overline{X}$ em várias situações. Obteremos resultados exatos ou aproximados, dependendo da distribuição de X na população. Quanto à função de probabilidade de S^2, voltaremos a ela em capítulos posteriores.

Consideremos, inicialmente, o caso de uma população Normal, isto é, a variável de interesse é $X \sim N(\mu, \sigma^2)$. Portanto, temos que $(X_1, X_2, ..., X_n)$ representa uma amostra aleatória cujos elementos são independentes e identicamente distribuídos, com densidade Normal de média μ e variância σ^2, ou seja,

$$X_i \sim N(\mu, \sigma^2), \qquad i = 1, ..., n;$$

$$X_i \text{ é independente de } X_j, \text{ para todo } i \neq j.$$

Vimos que, para quaisquer constantes $a_1, \ldots, a_n$, a combinação linear $\sum_{i=1}^{n} a_i X_i$ também tem distribuição de probabilidade dada pelo modelo Normal. A distribuição da média amostral segue diretamente deste resultado ao tomarmos $a_i = 1/n$, para $i = 1, ..., n$. Assim, $\overline{X} \sim N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$ e, com o auxílio das propriedades da esperança e variância apresentadas anteriormente, temos

$$\mu_{\overline{X}} = E(\overline{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n} n\mu = \mu\,;$$

$$\sigma^2_{\overline{X}} = Var(\overline{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}\,.$$

Logo, para uma coleção de variáveis aleatórias independentes com uma mesma distribuição de probabilidade, dada por um modelo Normal com média μ e

variância σ^2, a média amostral $\overline{X}$ também terá distribuição Normal, com média μ e variância σ^2/n.

Note que, pelo resultado acima, podemos concluir que à medida em que o tamanho da amostra cresce, a probabilidade de a média amostral estar na vizinhança da média populacional torna-se maior. Intuitivamente tal fato é esperado, uma vez que, ao aumentarmos o tamanho da amostra, estamos tendo mais informação e, de certa forma, fazendo a amostra ficar "mais parecida" com a população. É, assim, razoável acreditar que a média amostral será próxima da média populacional.

Exemplo 7.13: Considere uma amostra independente de tamanho n de uma variável $N(10, 16)$. Isto é, $X_1, X_2, \ldots, X_n$ são independentes e todas com distribuição Normal com média 10 e variância 16. Como se comporta $\overline{X}$ em função de n?

A variável aleatória $\overline{X}$ tem distribuição $N(10, 16/n)$ e o gráfico de sua densidade é apresentado, a seguir, para alguns valores de n.

Como podemos notar, à medida que n aumenta, a função densidade vai se concentrando ao redor da média 10, que é a média populacional, indicando maior probabilidade de amostras grandes fornecerem uma estimativa de $\overline{X}$ próxima da média populacional. □

Figura 7.1: Efeito de n na distribuição amostral de $\overline{X} \sim N(10, 16/n)$.

Exemplo 7.14: Suponha que a aceitação de um lote de 1000 peças ocorra apenas, se o comprimento médio de 10 peças, retiradas aleatoriamente do lote, estiver entre 5 e 10 cm. Sabe-se que o comprimento das peças é uma variável aleatória com distribuição Normal de média 7,5 cm e variância 20 cm². O que podemos dizer a respeito da aceitação do lote?

Se definirmos por X_i o comprimento da i-ésima peça retirada, $i = 1, \ldots, 10$, temos que a média das 10 peças a serem retiradas, representada por $\overline{X}$, terá distribuição Normal com média 7,5 cm e variância $20/10 = 2$ cm². Logo, a probabilidade de aceitarmos o lote será

$$P(5 < \overline{X} < 10) = P\left(\frac{5 - 7{,}5}{\sqrt{2}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{10 - 7{,}5}{\sqrt{2}}\right)$$
$$= P(-1{,}77 < Z < 1{,}77) = 0{,}9232\,;$$

obtida da tabela da $N(0,1)$. □

No que foi discutido até aqui, consideramos a distribuição amostral da média $\overline{X}$, calculada em uma amostra cujos elementos são constituídos por variáveis aleatórias independentes e com distribuição Normal. Na prática, muitas vezes não temos informações a respeito da distribuição das variáveis constituintes da amostra, o que nos impede de utilizar o resultado apresentado. Felizmente, satisfeitas certas condições, pode ser mostrado que, *para um tamanho de amostra suficientemente grande,* a distribuição de probabilidade da média amostral pode ser aproximada por uma distribuição Normal. Este fato é um dos teoremas mais importantes da área de Estatística e Probabilidade e é denominado *Teorema Central do Limite*, cuja demonstração requer técnicas mais avançadas e será omitida.

Teorema Central do Limite

Suponha uma amostra aleatória simples de tamanho n retirada de uma população com média μ e variância σ^2 (note que o modelo da variável aleatória não é especificado). Representando tal amostra por n variáveis aleatórias independentes $(X_1, \ldots, X_n)$ e, denotando sua média por $\overline{X}$, temos que

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{n\to\infty}{\longrightarrow} Z,$$

com $Z \sim N(0,1)$. □

Em palavras, o teorema garante que para n grande a distribuição da média amostral, devidamente padronizada, se comporta segundo um modelo Normal com média 0 e variância 1. De imediato, podemos notar a importância do Teorema Central do Limite, pois em muitas situações práticas, em que o interesse reside na média amostral, o teorema permite que utilizemos a distribuição Normal para estudar $\overline{X}$ probabilisticamente. Pelo teorema temos que quanto maior o tamanho da amostra, melhor é a aproximação. Estudos, envolvendo simulações, mostram que, em muitos casos, valores de n ao redor de 30 fornecem aproximações bastante boas para as aplicações práticas. Em casos em que a verdadeira distribuição dos dados é simétrica, excelentes aproximações são obtidas, mesmo com valores de n inferiores a 30.

Para verificar o efeito do tamanho da amostra sobre a distribuição de $\overline{X}$, vamos considerar diversos modelos de variáveis aleatórias e vários tamanhos de amostra. Com o auxílio do computador, simulamos a coleta de amostras de um determinado tamanho do modelo escolhido. Repetindo essa coleta um número grande de vezes e calculando as correspondentes médias amostrais, podemos obter um histograma dessas realizações, que ficaria muito próximo da função de probabilidade de $\overline{X}$. Por exemplo, fixe um tamanho da amostra e repita a coleta 100 vezes. Como cada amostra fornece uma média amostral, temos 100 médias amostrais observadas e com elas construímos um histograma. É claro que, quanto maior for a coleta e as repetições, mais aproximado será o histograma, da densidade de $\overline{X}$. Teremos, então, através dessa simulação, uma idéia de como $\overline{X}$ se comportaria numa amostra grande e poderemos perceber sua semelhança com a distribuição Normal, conforme assegura o Teorema Central do Limite.

Na Figura 7.2, apresentamos uma aplicação do procedimento descrito acima. Procuramos escolher modelos bem diferentes de modo a ilustrar a rapidez, no sentido do tamanho da amostra, e a qualidade da aproximação. Os modelos escolhidos foram Uniforme Discreto $(1, 10)$, Binomial $(n = 5,\ p = 0{,}2)$, Exponencial $(\lambda = 2)$ e o modelo contínuo definido pela densidade:

$$f(x) = \begin{cases} 1/8 & \text{se } 0 < x < 4\,; \\ 1/2 & \text{se } 4 \le x < 5\,. \end{cases}$$

Pode-se observar que, mesmo partindo de distribuição assimétricas, discretas ou contínuas, à medida em que o tamanho da amostra cresce, a distribuição de $\overline{X}$ vai se aproximando para a forma de um modelo Normal. A velocidade da convergência depende da distribuição inicial, sendo mais rápida nas distribuições simétricas.

Figura 7.2: Efeito do tamanho da amostra sobre a distribuição de $\overline{X}$.

Exemplo 7.15: Uma variável aleatória X assume os valores 3, 6 e 8 com, respectivamente, probabilidades 0,4; 0,3 e 0,3. Uma amostra com 40 observações é sorteada. A variável X não tem distribuição Normal e obtemos $\mu = 5{,}4$ e $\sigma^2 = 4{,}44$. Apesar de não ser simétrica, consideramos que 40 observações é uma amostra grande o suficiente para usar o Teorema Central do Limite. Para calcular a probabilidade da média amostral superar o valor 5, temos:

$$P(\overline{X} > 5) = P(\frac{\overline{X} - 5{,}4}{\sqrt{4{,}44/40}} > \frac{5 - 5{,}4}{\sqrt{4{,}44/40}}) \simeq P(Z > -1{,}20) = 0{,}8849;$$

com este último resultado obtido da tabela da $N(0,1)$. □

Exemplo 7.16: Em uma certa cidade, a duração de conversas telefônicas em minutos, originárias de telefones públicos, segue um modelo Exponencial com

parâmetro 1/3. Observando-se uma amostra aleatória de 50 dessas chamadas, qual será a probabilidade delas, em média, não ultrapassarem 4 minutos?

Representando por X a duração das chamadas, temos $X \sim Exp(1/3)$. Do Capítulo 6, temos que $E(X) = 3$ e $Var(X) = 9$. Admitindo uma amostra suficientemente grande, podemos calcular a probabilidade desejada da seguinte forma:

$$P(\overline{X} \leq 4) = P(\frac{\overline{X} - 3}{\sqrt{9/50}} \leq \frac{4 - 3}{\sqrt{9/50}}) \simeq P(Z \leq 2{,}36) = 0{,}9909.$$

Tendo em vista o alto valor de probabilidade encontrado, podemos dizer que é praticamente certo que a média amostral estará abaixo de 4 minutos. □

Uma aplicação importante do Teorema Central do Limite relaciona-se com a distribuição da proporção amostral. Recorde que definimos a proporção amostral como a fração dos indivíduos com uma dada característica em uma amostra de tamanho n, isto é,

$$\widehat{p} = \frac{\textit{número de indiv. na amostra com dada característica}}{n}.$$

Se construirmos para o i-ésimo indivíduo uma variável aleatória Y_i tal que

$$Y_i = \begin{cases} 1, & \text{se o indivíduo apresenta a característica;} \\ 0, & \text{caso contrário;} \end{cases}$$

podemos reescrever a proporção amostral como

$$\widehat{p} = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \sum_{i=1}^{n} \frac{Y_i}{n} = \overline{Y}.$$

Logo, a proporção amostral nada mais é do que a média de variáveis aleatórias convenientemente definidas. Assumindo que a proporção de indivíduos com a dada característica na população é p e que os indivíduos são selecionados aleatoriamente, temos que $Y_1, \ldots, Y_n$ formam uma seqüência de variáveis aleatórias independentes com distribuição de Bernoulli. Assim, $E(Y_i) = p$ e $Var(Y_i) = p(1-p)$. Logo,

$$E(\widehat{p}) = E(\sum_{i=1}^{n} \frac{Y_i}{n}) = p \qquad \text{e} \qquad Var(\widehat{p}) = Var(\sum_{i=1}^{n} \frac{Y_i}{n}) = \frac{p(1-p)}{n}.$$

Assim, $\widehat{p}$ é um estimador não viciado e consistente para p.

Tendo em vista o Teorema Central do Limite temos que para n suficientemente grande,

$$\frac{\overline{Y} - E(\overline{Y})}{\sqrt{Var(\overline{Y})}} = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{n\to\infty} N(0,1).$$

Notamos que a soma das variáveis aleatórias Y_i, digamos $W = \sum_{i=1}^{n} Y_i$, tem distribuição exata dada por uma Binomial com parâmetros n e p. Desta forma, probabilidades, envolvendo a proporção amostral, podem ser calculadas de modo exato usando essa distribuição. Caso o valor de n seja muito grande, essas probabilidades darão algum trabalho para serem calculadas e torna-se conveniente utilizar a aproximação Normal.

Exemplo 7.17: Suponha que a proporção de peças fora de especificação em um lote é de 40%. Tomada uma amostra de tamanho 30, a probabilidade desta amostra fornecer uma proporção de peças defeituosas menor que 0,50 pode ser calculada de forma exata pela Binomial e aproximada pelo modelo Normal.

Seja W a variável aleatória representando o número de peças defeituosas na amostra. Claramente, $W \sim b(30; 0{,}40)$. Logo, se $\widehat{p}$ representa a proporção amostral de peças defeituosas, temos que

$$\begin{aligned} P(\widehat{p} < 0{,}50) &= P(W/30 < 0{,}50) = P(W < 15) \\ &= \sum_{i=0}^{14} \binom{30}{i} 0{,}40^i\, 0{,}60^{30-i} = 0{,}8250. \end{aligned}$$

Considerando agora a aproximação Normal, temos, como conseqüência do Teorema Central do Limite

$$\widehat{p} \sim N(0{,}40, \frac{0,40(1-0,40)}{30}).$$

Assim,

$$P(\widehat{p} < 0{,}50) \simeq P(\frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0{,}50 - 0{,}40}{\sqrt{\frac{0{,}40(1-0{,}40)}{30}}}) = P(Z < 1{,}12) = 0{,}8686;$$

temos, então, mesmo para uma amostra não muito grande, uma proximidade razoável entre as duas respostas. □

Exercícios da Seção 7.3:

1. Uma variável de Bernoulli com probabilidade de sucesso p é amostrada, de forma independente, duas vezes. Apresente a função de probabilidade da média amostral.

2. O número de divórcios por indivíduo adulto casado, em certa comunidade, foi modelado pela variável aleatória D, cuja função de probabilidade é apresentada a seguir:

D	0	1	2	3
p_i	0,5	0,4	0,05	0,05

Uma amostra, representada por (D_1, D_2), foi sorteada com dois desses indivíduos e os seguintes estimadores, para a média de divórcios, foram considerados: $\widehat{\mu}_1 = \sqrt{D_1 D_2}$ e $\widehat{\mu}_2 =$ máximo $-$ mínimo. Para cada estimador, obtenha sua distribuição de probabilidade e verifique se é viciado.

3. Uma variável aleatória assume quatro valores $(-2, -1, 1, 2)$ com igual probabilidade. Para uma amostra de tamanho dois, obtenha a distribuição de S^2 e verifique se ele é não viesado para estimar a variância da variável.

4. Coleta-se uma amostra de 10 observações independentes de uma $N(2, 2)$. Determine a probabilidade de a média amostral:
   a. Ser inferior a 1.
   b. Ser superior a 2,5.
   c. Estar entre 0 e 2.

5. Supõe-se que o consumo mensal de água por residência em um certo bairro paulistano tem distribuição Normal com média 10 e desvio padrão 2 (em m^3). Para uma amostra de 25 dessas residências, qual é a probabilidade de a média amostral não se afastar da verdadeira média por mais de 1 m^3?

6. Um fabricante afirma que sua vacina contra gripe imuniza em 80% dos casos. Uma amostra de 25 indivíduos que tomaram a vacina foi sorteada e testes foram feitos para verificar a imunização ou não desses indivíduos. Se o fabricante estiver correto, qual é a probabilidade da proporção de imunizados na amostra ser inferior à 0,75? E superior à 0,85?

7. A resistência de vigas de madeira utilizadas na construção está sendo estudada. O fornecedor atesta que, em média, cada viga resiste a 3 toneladas com desvio padrão de aproximadamente 2 toneladas. Vinte dessas vigas serão sorteadas para serem utilizadas numa obra. Considerando que é verdadeira a informação do fornecedor e supondo que o modelo Normal é adequado, pergunta-se:

a. Qual a probabilidade de uma dessas vigas suportar menos do que 1 tonelada?

b. Qual a probabilidade de as vinte vigas suportarem, em média, pelo menos 2,5 toneladas?

c. Qual a probabilidade em (b), considerando agora 40 vigas e sem fazer a suposição de normalidade para os dados.

7.4 Estimação por Intervalo

Os estimadores discutidos até aqui são estimadores *pontuais*, pois fornecem como estimativa um único valor numérico para o parâmetro de interesse. Por serem variáveis aleatórias, os estimadores possuem uma distribuição de probabilidade e, levando este fato em consideração, podemos apresentar uma estimativa mais informativa para o parâmetro de interesse que inclua uma medida de precisão do valor obtido. Esse método de estimação, denominado *intervalo de confiança,* incorpora, à estimativa pontual do parâmetro, informações a respeito de sua variabilidade. Intervalos de confiança são obtidos através da distribuição amostral de seus estimadores.

Consideremos, inicialmente, o intervalo de confiança para a média μ de uma certa população Normal, com variância conhecida σ^2. Supondo uma amostra de tamanho n dada por $(X_1, \ldots, X_n)$, vimos que a média amostral tem distribuição Normal com a mesma média μ e variância σ^2/n. Assim,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Fixado um valor γ tal que $0 < \gamma < 1$, podemos encontrar um valor $z_{\gamma/2}$ tal que

$$P(|Z| < z_{\gamma/2}) = P(-z_{\gamma/2} < Z < z_{\gamma/2}) = \gamma.$$

O índice de $z_{\gamma/2}$ apresenta o valor de γ dividido por 2 uma vez que a "massa" γ deve ser distribuída igualmente em torno de 0 (veja a figura a seguir).

O valor $z_{\gamma/2}$ pode ser obtido da tabela da Normal padrão, localizando o valor de $\gamma/2$ no corpo da tabela e obtendo o valor $z_{\gamma/2}$ nas margens correspondentes. Feito isso, temos o intervalo

$$-z_{\gamma/2} < Z < z_{\gamma/2} \Rightarrow -z_{\gamma/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\gamma/2}$$

que pode ser reescrito como

$$\overline{X} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}}.$$

Assim, o intervalo de confiança para μ, com *coeficiente de confiança* γ, é dado por

$$\text{IC}(\mu, \gamma) = \left[\overline{X} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\,;\, \overline{X} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\right].$$

A interpretação do intervalo de confiança deve ser feita com cuidado. A expressão $\text{IC}(\mu, \gamma)$ envolve a quantidade $\overline{X}$ que é uma variável aleatória e, portanto, o intervalo obtido também é aleatório. A probabilidade que ele contenha o verdadeiro valor da média populacional μ é dada por γ. Ao coletar a amostra, $\overline{X}$ torna-se $\overline{x}_{obs}$ e, como conhecemos σ, n e $z_{\gamma/2}$, o intervalo passa a ser numérico. Desta forma, uma interpretação conveniente é a seguinte: *se obtivermos várias amostras de mesmo tamanho e, para cada uma delas, calcularmos os correspondentes intervalos de confiança com coeficiente de confiança* γ,

esperamos que a proporção de intervalos que contenham o valor de μ seja igual a γ. O exemplo a seguir ilustra os conceitos discutidos.

Exemplo 7.18: Suponha que os comprimentos de jacarés adultos de uma certa raça siga o modelo Normal com média μ desconhecida e variância igual a 0,01 m^2. Uma amostra de dez animais foi sorteada e forneceu média 1,69 m. Desejamos uma estimativa para o parâmetro desconhecido μ.

Uma vez que a distribuição de probabilidade de $\overline{X}$ é Normal com média μ e variância $\sigma^2/10 = 0{,}001\,\text{m}^2$, podemos proceder de forma semelhante ao desenvolvimento teórico apresentado acima e obter uma estimativa por intervalo para μ. Isto é, construimos um intervalo de confiança para μ com coeficiente γ. Estabelecendo $\gamma = 95\%$ obtemos da tabela da Normal $z_{\gamma/2} = z_{0,475} = 1{,}96$.

Segue, então, que

$$\begin{aligned} \text{IC}(\mu, 95\%) &= \left[1{,}69 - 1{,}96\sqrt{\frac{0{,}01}{10}}\,;\, 1{,}69 + 1{,}96\sqrt{\frac{0{,}01}{10}}\right] \\ &= [1{,}63;\, 1{,}75]. \end{aligned}$$

Adotando a interpretação, mencionada acima, de que em 100 intervalos construídos, 95 conteriam a verdadeira média, o que podemos dizer do intervalo obtido? De modo geral, o que é usualmente considerado é admitir que o intervalo de confiança calculado é um daqueles "bons", isto é, ele contém a verdadeira média μ. Essa razão explica a necessidade de, além de informar o intervalo numérico obtido, fornecer também o índice de confiança que foi utilizado. □

A *amplitude* do intervalo de confiança é dada pela diferença entre o extremo superior e inferior, isto é, $\overline{X} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}} - (\overline{X} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}}) = 2 \times z_{\gamma/2}\frac{\sigma}{\sqrt{n}}$, o que claramente indica que ela depende da confiança γ, do desvio padrão σ e do tamanho da amostra n. É usual se referir à *semi-amplitude*, como o *erro envolvido na estimação*.

Veremos, a seguir, como a expressão da amplitude é influenciada pelo seus termos e destacaremos os aspectos intuitivos dessa influência.

Cada vez que aumentamos a confiança, o valor de $z_{\gamma/2}$ fica maior e consequentemente a amplitude do intervalo aumenta. Isto é esperado, pois intervalos maiores tem maior possibilidade de "captura" do verdadeiro valor de μ. Que confiança você atribuiria ao intervalo de 0 a 2,5 metros para a altura média de indivíduos adultos de uma cidade?

A variabilidade da variável aleatória na população é também um fator importante. Uma variância grande indica a possibilidade de um considerável

distanciamento dos possíveis valores amostrais em relação à média populacional, cujo intervalo de confiança estamos obtendo. Dependendo do seu tamanho n, a amostra pode fornecer um valor médio ($\overline{x}_{obs}$) muito influenciado pelos valores extremos.

Com relação à amostra, temos uma clara intuição de que, quanto maior for seu tamanho, maior será a quantidade de informação disponível. Note que, pela expressão da amplitude, para uma mesma variabilidade σ e confiança γ, valores maiores de n produzem intervalos menores e, portanto, mais informativos. Por exemplo, para a altura média de indivíduos, o intervalo 0 a 2,5 metros é menos informativo do que o intervalo 1,3 a 1,7 metros.

Exemplo 7.19: A vida média de baterias automotivas de uma certa marca está sendo estudada. Baseado em estudos similares, com outras marcas, é possível admitir que a vida dessas baterias segue a distribuição Normal com desvio padrão de 4,5 meses. De qual tamanho deverá ser a amostra, para que a amplitude do intervalo de 90% de confiança para a vida média seja de 3 meses?

Para calcular o valor de n, consideramos a equação:

$$2 \times z_{\gamma/2}\frac{\sigma}{\sqrt{n}} = 3.$$

Com os valores de $z_{\gamma/2} = 1{,}64$ ($\gamma = 90\%$) e $\sigma = 4{,}5$ temos

$$\sqrt{n} = \frac{2\, z_{\gamma/2}\, \sigma}{3} = \frac{2 \times 1{,}64 \times 4{,}5}{3} = 4{,}92.$$

Como o valor de n precisa ser um número inteiro, escolhemos o maior inteiro que contém $(4{,}92)^2$, obtendo $n = 25$. Dessa forma, a amplitude do intervalo a ser construído será ligeiramente menor do que 3 e, portanto, o intervalo será mais informativo. □

A aplicação do Teorema Central do Limite permite a obtenção de intervalos de confiança para μ, quando a distribuição das variáveis aleatórias, que constituem a amostra, não segue um modelo Normal. Neste caso, o intervalo construído terá um coeficiente de confiança *aproximadamente* igual a γ, sendo que esta aproximação melhora à medida que aumenta o tamanho da amostra.

Exemplo 7.20: Um provedor de acesso à Internet está monitorando a duração do tempo das conexões de seus clientes, com o objetivo de dimensionar seus equipamentos. São desconhecidas a média e a distribuição de probabilidade desse tempo, mas o desvio padrão, por analogia a outros serviços, é considerado igual a

$\sqrt{50}$ minutos. Uma amostra de 500 conexões resultou num valor médio observado de 25 minutos. O que dizer da verdadeira média, com confiança 92%?

O tempo de duração das conexões não tem distribuição conhecida e a falta dessa informação precisará ser contornada. Uma alternativa viável é buscar a aplicação do Teorema Central do Limite. O tamanho da amostra é igual a 500 e, portanto, será bastante razoável aplicar esse teorema para utilizar a distribuição Normal. Na verdade, com esse tamanho de amostra, teremos uma aproximação bastante boa.

Nestes termos, um intervalo de confiança aproximado de 92%, para a média do tempo de conexão, será dado por

$$\begin{aligned} \mathrm{IC}(\mu, 92\%) &\simeq \left[\overline{X} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\,;\, \overline{X} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\right] \\ &\simeq \left[25 - 1{,}75\sqrt{\frac{50}{500}}\,;\, 25 + 1{,}75\sqrt{\frac{50}{500}}\right] \\ &\simeq [\,24{,}45\,;\,25{,}55\,]. \end{aligned}$$

Note que, em virtude do uso do Teorema Central do Limite, obtemos um intervalo com coeficiente de confiança *aproximadamente* 92%. □

Como outro exemplo de aplicação do Teorema Central do Limite, vamos considerar intervalos de confiança para a proporção p.

Exemplo 7.21: Pretende-se estimar a proporção p de cura, através do uso de um certo medicamento em doentes contaminados com cercária, que é uma das formas do verme da esquistossomose. Um experimento consistiu em aplicar o medicamento em 200 pacientes, escolhidos ao acaso, e observar que 160 deles foram curados. Que podemos dizer da proporção p na população em geral?

Uma estimativa pontual para p é dada por $\widehat{p}_{obs} = 160/200 = 0{,}8$. Entretanto, podemos dizer mais do que isso. Apesar da proporção amostral $\widehat{p}$ não ter distribuição Normal, o Teorema Central do Limite nos garante que, para um tamanho de amostra grande, poderemos aproximá-la para a Normal. Desse modo segue que

$$\widehat{p} \sim N(p, \frac{p(1-p)}{n}).$$

Assim, um intervalo de confiança com coeficiente aproximado $\gamma = 0{,}95$ (e portanto $z_{\gamma/2} = 1{,}96$) é dado por:

$$\begin{aligned} \mathrm{IC}(p, 95\%) &= \left[\widehat{p} - 1{,}96\sqrt{\frac{p(1-p)}{n}}\,;\, \widehat{p} + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right] \\ &= \left[0{,}8 - 1{,}96\sqrt{\frac{p(1-p)}{n}}\,;\, 0{,}8 + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right]. \end{aligned}$$

Note que neste caso, como p é desconhecido, o intervalo ainda não pode ser calculado diretamente pois envolve uma quantidade desconhecida dentro da raiz quadrada.

Uma possível solução é substituirmos $p(1-p)$ por $\widehat{p}_{obs}(1 - \widehat{p}_{obs})$. Dessa forma, estamos utilizando a estimativa pontual obtida no lugar do parâmetro desconhecido p. O intervalo será:

$$\begin{aligned} \mathrm{IC}_1(p, 95\%) &= \left[0{,}8 - 1{,}96\sqrt{\frac{0{,}8 \times 0{,}2}{200}}\,;\, 0{,}8 + 1{,}96\sqrt{\frac{0{,}8 \times 0{,}2}{200}}\right] \\ &= [\,0{,}745;\, 0{,}855\,]. \end{aligned}$$

Outra abordagem possível é baseada no fato que a expressão $p(1-p)$ tem valor máximo igual a $1/4$, quando $0 \leq p \leq 1$. Verifique essa afirmação, fazendo o gráfico da função $p(1-p)$ com p variando de 0 a 1. Nesse caso, podemos obter um intervalo de confiança substituindo $p(1-p)$ por $1/4$:

$$\begin{aligned} \mathrm{IC}_2(p, 95\%) &= \left[0{,}8 - 1{,}96\sqrt{\frac{1/4}{200}}\,;\, 0{,}8 + 1{,}96\sqrt{\frac{1/4}{200}}\right] \\ &= [\,0{,}731;\, 0{,}869]. \end{aligned}$$

Note que este intervalo tem amplitude maior que IC_1. □

Temos, portanto, duas alternativas para o cálculo de intervalos de confiança para p. A primeira, dada por IC_1, é usualmente denominada abordagem *otimista*, pois parte da crença que a estimativa obtida está suficientemente próxima de p de tal forma que a variância dada por $p(1-p)/n$ é bem aproximada por $\widehat{p}(1 - \widehat{p})/n$. Já a outra abordagem, calculada em IC_2, é conhecida como abordagem *conservativa*, pois preferimos substituir a variância por um valor seguramente maior do que o real. Assim, estamos nos assegurando que o coeficiente de confiança será de, no mínimo, γ. Lembrando que a variância de

um estimador é uma medida de sua precisão, ao utilizarmos a abordagem conservativa estamos aceitando uma menor precisão para $\widehat{p}$, o que se reflete numa maior amplitude do intervalo de confiança IC_2, quando comparado ao intervalo de confiança otimista.

Na Tabela 7.2, apresentamos um resumo dos intervalos construídos neste capítulo. Intervalos de confiança para a média, no caso da variância ser desconhecida, serão desenvolvidos no próximo capítulo, quando um novo modelo de densidade será definido.

Tabela 7.2: Intervalos de confiança para μ e p.

Parâmetro	Intervalo de Confiança
μ	$[\overline{X} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\,;\,\overline{X} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}}]$
p	$[\widehat{p} - z_{\gamma/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}};\widehat{p} + z_{\gamma/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}]$ (otimista)
p	$[\widehat{p} - z_{\gamma/2}\sqrt{\frac{1}{4n}};\widehat{p} + z_{\gamma/2}\sqrt{\frac{1}{4n}}]$ (conservativo)

Exercícios da Seção 7.4:

1. Por analogia a produtos similares, o tempo de reação de um novo medicamento pode ser considerado como tendo distribuição Normal com desvio padrão igual a 2 minutos (a média é desconhecida). Vinte pacientes foram sorteados, receberam o medicamento e tiveram seu tempo de reação anotado. Os dados foram os seguintes (em minutos): 2,9; 3,4; 3,5; 4,1; 4,6; 4,7; 4,5; 3,8; 5,3; 4,9; 4,8; 5,7; 5,8; 5,0; 3,4; 5,9; 6,3; 4,6; 5,5 e 6,2. Obtenha um intervalo de confiança para o tempo médio de reação. Use γ= 96%.

2. Uma amostra de 25 observações de uma Normal (μ, 16) foi coletada e forneceu uma média amostral de 8. Construa intervalos com confiança 80%, 85%, 90% e 95% para a média populacional. Comente as diferenças encontradas.

3. Será coletada uma amostra de uma população Normal com desvio padrão igual a 9. Para uma confiança de γ=90%, determine a amplitude do intervalo de confiança para a média populacional nos casos em que o tamanho da amostra é 30, 50 ou 100. Comente as diferenças.

4. Uma amostra em 100 cidades brasileiras, de até 20 mil habitantes, indicou que o valor médio da hora aula para os professores do ensino fundamental em escolas municipais é de R$ 2,5. Obtenha um intervalo de confiança para o valor médio nacional da hora aula em cidades do tipo mencionado. Baseado em estudos anteriores, o desvio padrão é assumido ser igual a R$ 1,1. Use $\gamma = 0{,}95$.

5. Numa pesquisa com 50 eleitores, o candidato José João obteve 0,34 da preferência dos eleitores. Construa, para a confiança 94%, os intervalos otimista e conservador de confiança para a proporção de votos a serem recebidos pelo candidato mencionado, supondo que a eleição fosse nesse momento.

7.5 Exercícios

1. Foram sorteadas 20 escolas de ensino fundamental da rede privada, na cidade de São Paulo e observado o número classes de 1a. série em cada uma delas. Os resultados foram: 2, 3, 3, 4, 3, 2, 1, 2, 3, 2, 3, 4, 5, 4, 2, 4, 5, 5, 1, e 2. Deseja-se estimar o número médio de classes nesse tipo de escola, com vistas a um futuro levantamento de disponibilidade de vagas. Obtenha as estimativas correspondentes aos seguintes estimadores propostos:

$$\widehat{\mu}_1 = \text{mediana amostral};$$
$$\widehat{\mu}_2 = \text{moda amostral};$$
$$\widehat{\mu}_3 = \overline{X}\,.$$

Tendo em vista o objetivo pretendido, discuta as vantagens de cada um deles.

2. O Conselho Regional de Odontologia recomenda visitas periódicas ao dentista e, para orientar sua campanha de divulgação, realizou uma pesquisa com 100 crianças com idades de 12 a 14 anos. Quanto ao número de visitas no último ano, a amostra resultou em uma média de 0,5 e mediana e moda iguais a 0. Com base nesses dados comente as afirmações abaixo.

a. A maioria não visitou o dentista no último ano.

b. Metade da população dessas crianças nunca foi ao dentista.

c. No último 0,5 ano as crianças tiveram em média 1 visita.

d. Talvez algumas crianças tenham feito mais de uma visita no último ano.

3. Um grupo de 15 alunos do curso de Veterinária foi sorteado e respondeu a uma pesquisa sobre o número de vestibulares antes de conseguirem entrar na universidade. Os resultados foram: 3, 1, 2, 1, 2, 3, 3, 2, 3, 3, 1, 1, 3, 2 e 3. Calcule as estimativas para a variância segundo os estimadores:

$$\widehat{\sigma}_1^2 = (\text{mediana} - \text{mínimo})^2\,;$$
$$\widehat{\sigma}_2^2 = (\text{máximo} - \text{mediana})^2\,;$$
$$\widehat{\sigma}_3^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

4. Uma variável aleatória X assume os valores -1, 0 e 1 com igual probabilidade. Para uma amostra de tamanho 2, obtenha a distribuição de probabilidade de $\overline{X}$ e verifique que ele não é viciado para estimar a média de X.

5. Estatísticas do Departamento de Trânsito sobre o envolvimento em acidentes, de motoristas com pouca experiência (até 2 anos de habilitação), indicam que o seguinte modelo pode ser adotado:

Número de Acidentes (X)	0	1	2	3	4
p_i	0,3	0,3	0,2	0,1	0,1

a. Determine $E(X)$ e $Var(X)$.

Para uma amostra aleatória de 2 desses "jovens" motoristas, obtenha:

b. A distribuição amostral de $\overline{X}$.

c. $E(\overline{X})$ e $Var(\overline{X})$.

d. Os histogramas de X e $\overline{X}$ (num mesmo diagrama).

6. Determine a probabilidade de a média amostral de 3 observações aleatórias de uma $b(2;\,0{,}3)$ ser inferior à 1.

7. Sendo X uma Bernoulli com $p = 0{,}6$ determine, para uma amostra de tamanho 3, a função de probabilidade da mediana amostral.

8. Uma amostra de duas observações da variável X é retirada. Nos vários casos listados abaixo, determine a função de probabilidade de S^2 e constate se ele é não viciado para estimar a variância.

a. X é Uniforme Discreta em $\{1, 2, 3\}$.

b. X é Bernoulli com $p = 0{,}2$.

c. X é Binomial com $n = 3$ e $p = 0{,}5$.

9. Para uma Normal (5, 10) coletou-se uma amostra de tamanho 25. Calcule:

a. $P(\overline{X} \leq 4{,}8)$.

b. $P(4{,}5 \leq \overline{X} \leq 5{,}3)$.

c. $P(\overline{X} \leq 4{,}7 \text{ ou } \overline{X} \geq 5{,}1)$.

10. Em 10 observações de uma variável seguindo o modelo Normal com média 3 e desvio padrão 2, qual será a probabilidade de a média amostral:

a. Ser superior a 1,5?

b. Ser inferior a 0?

c. Não se afastar da verdadeira média por mais de 1 unidade?

11. Trinta observações de uma Normal com média μ e variância 36 são coletadas.

a. Calcule $P(|\overline{X} - \mu| \leq 3)$.

b. Determine o valor de a tal que $P(|\overline{X} - \mu| \geq a) \simeq 0{,}9$.

12. Sendo a variável amostrada uma Normal de média μ e variância 25, obtenha o valor de $P(|\overline{X} - \mu| \leq 2)$ nos casos de tamanho da amostra igual a 2, 20 e 60. Comente os resultados obtidos.

13. Considere uma amostra de tamanho 30 de uma população Normal de média μ e variância σ^2. Determine $P(|\overline{X} - \mu| \leq 1)$ nos casos em que σ^2 é igual a 16, 64 e 100. Qual a conclusão?

14. A duração do "tonner" de uma máquina de fotocópias pode ser modelado como Normal com média 15 e desvio padrão 2 (em milhares de cópias). Para uma amostra de 12 fotocopiadoras a duração do "tonner" será observada e pergunta-se a probabilidade de, em média, durar:

a. Menos de 16 mil cópias?

b. Mais de 13 mil cópias?

c. Entre 12 e 14 mil cópias?

15. Uma máquina enche pacotes de café com um peso que se comporta como uma variável aleatória Normal de média 200 gramas e desvio padrão 10 gramas. Uma amostra de 25 pacotes é sorteada e pergunta-se:

a. Qual é o número esperado de pacotes da amostra com peso inferior a 205 gramas?

b. Qual é a probabilidade de que o peso total dos pacotes da amostra não exceda 5125 gramas?

16. Para se ajustar a uma máquina, a correia deve ter entre 60 e 62 cm de comprimento. Tendo em vista o processo de fabricação, o comprimento dessas correias pode ser considerado como uma variável aleatória com distribuição Normal de média 60,7 cm e desvio padrão 0,8 cm. Pergunta-se:

a. Qual a probabilidade de uma correia, escolhida ao acaso, poder ser usada na máquina?

b. Um grande revendedor dessas correias estabelece um controle de qualidade nos lotes que compra da fábrica: ele sorteia 4 correias do lote e só aceita o lote se o comprimento médio estiver dentro do tamanho aceito pela máquina. Calcule a probabilidade de aceitação do lote.

17. Desejamos coletar uma amostra de uma variável aleatória X com distribuição Normal de média desconhecida e variância 30. Qual deve ser o tamanho da amostra para que, com 0,92 de probabilidade, a média amostral não difira da média da população por mais de 3 unidades?

18. Seja $X \sim N(\mu, 36)$.

a. Para uma amostra de tamanho 50, obtivemos média amostral 18,5. Construa intervalos com confiança 91%, 96% e 99% para μ.

b. Para uma confiança de 94%, construa intervalos de confiança supondo três tamanhos de amostra 25, 50 e 100 (admita que todos forneceram a mesma média amostral igual a 18,5).

c. Comente sobre a precisão dos intervalos construídos em (a) e (b).

19. Interprete e comente as afirmações abaixo:

a. "A média de salário inicial para recém formados em Economia está entre 7 e 9 salários mínimos com confiança 95%".

b. "Quanto maior for o tamanho da amostra, maior é a probabilidade da média amostral estar próxima da verdadeira média".

20. Num grupo de pacientes, o nível de colesterol é uma variável aleatória com distribuição Normal, de média desconhecida e variância 64 $(mg/ml)^2$.

a. Para uma amostra de 46 indivíduos que forneceu nível médio de colesterol de 120 mg/ml, construa o intervalo de confiança de 88%.

b. Se você desejasse diminuir a amplitude do intervalo encontrado em a), quais seriam suas alternativas?

21. O consumo de combustível é uma variável aleatória com parâmetros dependendo do tipo de veículo. Suponha que, para um certo automóvel, o desvio padrão do consumo seja conhecido e igual a 2 km/l. Porém precisamos informações sobre o consumo médio. Para tal, coletamos uma amostra de 40 automóveis desse modelo e observamos o seu consumo.

a. Quem seria um estimador do consumo médio para todos os automóveis desse tipo?

b. Se a amostra forneceu um consumo médio de 9,3 km/l, construa um intervalo de confiança (94%) para a média de consumo desses carros.

c. Se a amplitude de um intervalo de confiança, construído a partir dessa amostra, é de 1,5; qual teria sido o coeficiente de confiança?

22. O intervalo [35,21; 35,99], com confiança 95% foi construído a partir de uma amostra de tamanho 100, para a média μ de uma população Normal com desvio padrão igual a 2.

a. Qual o valor encontrado para a média dessa amostra?

b. Se utilizássemos essa mesma amostra, mas uma confiança de 90%, qual seria o novo intervalo de confiança?

23. A dosagem de certa substância no sangue segue distribuição Normal com média μ e desvio padrão 15 mg/l. Se uma amostra de tamanho 25 for coletada, determine:

a. A probabilidade de $|\overline{X} - \mu|$ ser inferior a 5.

b. O intervalo para μ com confiança 98%, se temos $\overline{x}_{obs}$= 98 mg/l.

24. Uma amostra de trinta dias do número de ocorrências policiais em um certo bairro de São Paulo, apresentou os seguintes resultados: 7, 11, 8, 9, 10, 14, 6, 8, 8, 7, 8, 10, 10, 14, 12, 14, 12, 9, 11, 13, 13, 8, 6, 8, 13, 10, 14, 5, 14 e 10.

a. Fazendo as suposições devidas, construa um intervalo de confiança para a proporção de dias *violentos* (com pelo menos 12 ocorrências). Use os dois enfoques e a confiança de 88%.

b. Em um ano (360 dias) e com a mesma confiança de 88%, qual seria a estimativa do número de dias violentos nesse bairro?

c. Dê uma interpretação para os intervalos encontrados em (a).

25. Antes de uma eleição , um determinado partido está interessado em estimar a probabilidade p de eleitores favoráveis ao seu candidato. Uma amostra piloto de tamanho 100 revelou que 60% dos eleitores eram favoráveis ao candidato.

a. Utilizando a informação da amostra piloto, determine o tamanho da amostra para que, com 0,8 de probabilidade, o erro cometido na estimação seja no máximo 0,05.

b. Se na amostra final, com o tamanho obtido em (a), observou-se que 51% dos eleitores eram favoráveis ao candidato, construa um intervalo de confiança para p, com confiança 95%.

26. A análise de ocorrência de um mineral numa região é uma variável aleatória com média 4 e variância 3/2. A unidade de medida é porcentagem de mineral por unidade de volume. Para uma amostra de tamanho 20:

a. Que dizer da distribuição de $\overline{X}$?

b. Que tamanho deveria ter a amostra para que $P(3{,}5 < \overline{X} \leq 4{,}5) = 0{,}95$?

27. O tempo de reação de uma pessoa a certa droga é considerado uma variável aleatória com média 5 minutos e desvio padrão 3 minutos. Esse tempo foi medido em uma amostra de 80 pessoas escolhidas, sem reposição, na cidade de São Paulo. Pergunta-se a probabilidade de:

a. O tempo médio amostral ser inferior a 5,5?

b. O tempo médio na amostra não diferir da verdadeira média por mais de 0,4?

28. O comprimento de certo tipo de eixo, produzido pela empresa *Duroaço*, tem uma pequena variação de peça para peça. A lei de probabilidade, seguida por esses comprimentos, é desconhecida, porém admite-se que o desvio padrão é 4 milímetros. Uma amostra aleatória de 100 desses eixos, forneceu um comprimento médio de 4,52 milímetros.

a. Construa um intervalo, com confiança 90%, para a média do comprimento desses eixos fabricados pela *Duroaço*.

b. Dê uma interpretação para o intervalo encontrado. Será que podemos dizer que o intervalo encontrado em (a) tem probabilidade de 0,90 de conter a verdadeira média?

29. Numa pesquisa de mercado, desejamos estimar a proporção de pessoas que compram o sabonete *Bom-cheiro*.

a. Que tamanho de amostra devemos colher para que, com probabilidade 0,9; a estimativa não se desvie do verdadeiro valor por mais de 0,05?

b. Se tivermos a informação adicional de que a aceitação do sabonete *Bom-cheiro* é no mínimo 0,8, qual deve ser então o tamanho da amostra?

c. Decidimos colher uma amostra de tamanho 81. Qual o erro máximo que cometemos com probabilidade 0,9?

d. Para a amostra de tamanho 81, qual a probabilidade de que o erro máximo seja 0,08?

30. Considere a variável X com distribuição Bernoulli de parâmetro p. Uma amostra aleatória de tamanho 2 é retirada com o objetivo de estimar a média de X. Dois estimadores são propostos: $\hat{\mu}_1 = \overline{X}$ e $\hat{\mu}_2 = 0{,}8\,X_1 + 0{,}2\,X_2$. Obtenha a distribuição amostral desses estimadores em função de p e discuta suas propriedades.

31. Sendo $X \sim b(n = 10, p = 0{,}5)$, pergunta-se:

a. Para uma amostra de 2 observações dessa variável qual é a probabilidade da média amostral ser superior a 9? Justifique.

b. Para uma amostra de 100 observações dessa variável qual é a probabilidade da média amostral ser superior a 4,7? Justifique.

32. Para estimar a média das alturas (em metros) numa certa população, dois institutos de pesquisa coletaram cada um a sua amostra e usaram estimadores diferentes. Os resultados estão na tabela abaixo:

	Tamanho	Estimador	Valor Observado
Instituto 1	$n_1 = 100$	$\widehat{\mu}_1 = \overline{X}$	1,68
Instituto 2	$n_2 = 200$	$\widehat{\mu}_2 = (max + min)/2$	1,73

Apresente justificativas ao responder as questões abaixo:

a. Você acha que o valor 1,73 está mais perto da verdadeira média por ter vindo de uma amostra maior?

b. A verdadeira média deve estar no intervalo 1,68 até 1,73 ?

c. Indique qual das estimativas você preferiria usar.

33. O tempo de emissão de extratos, em segundos, pelo caixa eletrônico de um banco foi modelado segundo uma distribuição Exponencial com parâmetro 1/40. Para uma amostra aleatória de 50 clientes que solicitaram extratos:

a. Qual a probabilidade do segundo cliente sorteado na amostra demorar mais de 30 segundos na sua solicitação?

b. Determine a probabilidade de que o intervalo médio de emissão, entre os clientes amostrados, seja inferior a 35 segundos?

34. O tempo de espera, em minutos, na fila de votação numa certa zona eleitoral com urna eletrônica, foi modelado segundo uma distribuição Uniforme Contínua com valores entre 0 e 30. Para uma amostra aleatória de 100 eleitores, responda:

a. Qual a probabilidade do último eleitor na amostra demorar mais de 20 minutos?

b. Qual a probabilidade da média da amostra ser inferior a 18 minutos?

c. Você deseja pedir a um amigo que espere um tempo t para lhe dar uma carona. Usando a média da amostra, qual deve ser o valor de t para não perder a carona com probabilidade 0,8?

35. Admita que o número de viagens ao exterior é uma variável aleatória, com a distribuição abaixo, sendo que o valor de θ depende da profissão exercida.

Núm. de viagens	$\theta - 2$	$\theta - 1$	θ	$\theta + 1$	$\theta + 2$
Probab.	0,1	0,2	0,2	0,4	0,1

$\theta \in \{2, 3, 4, \ldots\}$.

Um indivíduo, sorteado aleatoriamente, indicou que fez 4 viagens ao exterior. Com base nessa informação, responda as questões abaixo:

a. Quais são os valores possíveis de θ para a profissão do indivíduo escolhido?

b. Dê uma estimativa para o valor de θ. Indique o critério utilizado.

36. (Use o computador) Com o objetivo de *simular* a distribuição amostral de $\bar{X}$, realize as seguintes tarefas:

a. Gere 100 amostras de tamanho 30 de uma Normal com média 200 e desvio-padrão 5. Calcule, então, a média de cada amostra e faça o histograma correspondente à esse conjunto de médias amostrais. Qual conclusão pode ser tirada? Calcule medidas descritivas das 100 médias obtidas e comente.

b. Repita o item (a) para as amostras geradas de uma Binomial com parâmetros $n = 50$ e $p = 0{,}45$. Comente a respeito do histograma e das medidas descritivas obtidas, tendo em vista o Teorema Central do Limite.

37. (Use o computador) Para estudar o comportamento do estimador S^2, gere 120 amostras de tamanho 20 de uma Exponencial de parâmetro 1/5. Em cada amostra, calcule a estimativa de S^2. Com as 120 estimativas obtidas, utilize as medidas descritivas e o histograma para comparar esses resultados com a variância do modelo Exponencial.

38. (Use o computador) Considere os dados apresentados no arquivo *aeusp.txt* (ver Exercício 26 no Capítulo 1) que representam uma amostra das comunidades de baixa renda do Butantã- SP. Com essas informações, obtenha estimativas, da média e da variância dessa população para as seguintes variáveis: Idade, Temposp, Resid e Renda.

Os exercícios, a seguir, têm motivação didática e não prática, pois conhecendo a verdadeira média não faz sentido calcular intervalos de confiança para ela.

39. (Use o computador) Simule a coleta de 80 observações de uma variável Normal$(10, 25)$ e obtenha o intervalo de confiança (92%) para a média. Repita esse procedimento 100 vezes, isto é, obtenha 100 intervalos de confiança. Verifique quantos deles contêm a verdadeira média.

40. (Use o computador) Gere 200 observações do modelo Uniforme Contínuo $[0, 20]$. Construa um intervalo de confiança 95% para a média. Repetindo 120 vezes esse procedimento, quantos intervalos conterão a verdadeira média? Comente os resultados e as suposições feitas.

Capítulo 8

Inferência Estatística - Testes de Hipóteses

8.1 Introdução

Apresentaremos, neste capítulo, um dos principais tópicos da Inferência Estatística conhecido como teste de hipóteses. Motivamos conceitualmente a técnica através de dois exemplos que ilustram sua utilidade e o tipo de questões que procura responder.

Exemplo 8.1: Suponha que, entre pessoas sadias, a concentração de certa substância no sangue se comporta segundo um modelo Normal com média 14 unidades/ml e desvio padrão 6 unidades/ml. Pessoas sofrendo de uma doença específica têm a concentração média da substância alterada para 18 unidades/ml. Admitimos que o modelo Normal, com desvio padrão 6 unidades/ml, continua representando de forma adequada a concentração da substância em pessoas com a doença. A figura a seguir ilustra a situação descrita.

Note que as curvas, representando as concentrações, irão se cruzar em algum momento, fazendo com que uma certa proporção de indivíduos na população sadia possa apresentar valores de concentração tão altos quanto aqueles observados para pessoas doentes, ainda que este evento ocorra com baixa probabilidade.

Desejamos averiguar se um certo tratamento, proposto para combater a doença, é eficaz. Uma amostra aleatória de tamanho $n = 30$ é selecionada entre indivíduos doentes que foram submetidos ao tratamento. Representemos as concentrações dos indivíduos da amostra por $X_1, \ldots, X_{30}$. Sabemos que para $i = 1, 2, \ldots, 30$, temos $X_i \sim N(\mu, 36)$, sendo $\mu = 14$ ou $\mu = 18$ dependendo do tratamento ser eficiente ou não. Caso a amostra de 30 valores forneça valor médio de concentração alto e "próximo" de 18, teremos evidências de que o tratamento não é eficaz, ao passo que um valor baixo e "próximo" de 14 unidades/ml nos revelaria crer que o tratamento apresenta resultados satisfatórios. Isto é, se o tratamento for eficaz, então os 30 indivíduos podem ser vistos como membros da população com concentração modelada por uma $N(14, 36)$, caso contrário, eles pertencerão à população $N(18, 36)$. A caracterização do que significa ser "próximo" depende, entre outros fatores, da variabilidade da concentração na população. Como a amostra de 30 indivíduos é aleatória, precisaremos estudar probabilisticamente o problema, o que será feito através da técnica estatística denominada *teste de hipóteses para a média com variância conhecida* e será objeto de discussão na Seção 8.2. □

Exemplo 8.2: Deseja-se estudar a tolerância de um equipamento eletrônico ao número de impactos termo-elétricos. Pelas características de fabricação do equipamento, é possível admitir que a probabilidade de falha seja constante, isto é, após cada impacto, existe uma probabilidade p de que ele falhe. Representando por X a variável aleatória *número de impactos anteriores à falha*, pretende-se verificar se o modelo Geométrico com $p = 0{,}4$ é adequado para caracterizar essa variável.

A decisão que precisamos tomar é aceitar ou não o modelo sugerido. Note que não é um simples valor de parâmetro que está sendo testado, mas sim a adequação ou não de uma função de probabilidade.

Nossa conclusão será baseada nos valores amostrados da variável de interesse. Neste exemplo, foram escolhidos, ao acaso, 80 equipamentos para serem submetidos a sucessivos impactos termo-elétricos até a ocorrência da primeira falha. Isto é, temos 80 realizações da variável X que serão denotadas por $X_1, \ldots, X_{80}$. Dessa forma, X_i representa o número de impactos até a ocorrência de falha no i-ésimo equipamento testado, $i = 1, 2, \ldots, 80$. A partir dos valores observados, podemos construir uma tabela de freqüência indicando os valores (ou grupo de valores) e suas respectivas freqüências de ocorrências. Podemos criar, caso seja conveniente, um certo número de categorias para X contendo um ou mais valores da variável.

Se o modelo Geométrico fosse verdadeiro, deveríamos ter, para cada valor da variável X, a freqüência observada não muito diferente daquela que seria calculada através da distribuição Geométrica que é denominada *freqüência esperada*. Devido às aleatoriedades envolvidas, poderão ocorrer pequenos desvios e ainda assim o modelo ser adequado. Portanto, vamos decidir pela aceitação ou não da distribuição Geométrica, avaliando a "distância" entre as freqüências esperadas e observadas. Para avaliar essa distância, consideramos a quantidade aleatória não negativa Q^2 dada pela expressão

$$Q^2 = \sum_{i=1}^{k} \frac{(\mathrm{o}_i - \mathrm{e}_i)^2}{\mathrm{e}_i},$$

com k sendo o total de categorias e o_i e e_i, respectivamente, a freqüência observada na amostra e a freqüência esperada da categoria i. Note que valores pequenos de Q^2 indicam que o modelo pode ser adequado aos dados. Por outro lado, valores acima de um certo ponto crítico q_c devem levar à decisão de rejeitar o modelo. A determinação de q_c depende da distribuição de Q^2 e do possível erro associado à decisão. Voltaremos a esse problema na Seção 8.5. □

Os exemplos acima apresentam duas situações típicas de teste de hipóteses. No Exemplo 8.1, o teste se refere à média populacional, enquanto que no Exemplo 8.2, a decisão é a respeito da função de probabilidade. Apesar de estarem em diferentes contextos, o procedimento conceitual do teste é similar. Na próxima seção, vamos apresentar a estrutura geral dessa técnica.

Exercícios da Seção 8.1:

1. Identifique as hipóteses que estão sendo testadas em cada caso.

a. A companhia de transporte afirma que, em média, o intervalo entre sucessivos ônibus é de 15 minutos. Uma associação de usuários de transportes coletivos acha que a pontualidade é muito importante e pretende testar a afirmação da companhia. $\mu = 15$ contra $\mu \neq 15$

b. Os amortecedores de automóveis que circulam em cidades duram em média 30 mil quilômetros, segundo informação de algumas oficinas especializadas. Um proprietário de automóvel deseja testar essa afirmação.

c. Um veterinário conseguiu ganho médio diário de 3 litros de leite por vaca com uma nova composição de ração. Um pecuarista acredita que o ganho não é tão grande assim. $\mu = 3$ contra $\mu < 3$

2. Garrafas de cerveja deveriam conter 600 ml porém existem flutuações aleatórias. Os órgãos de fiscalização permitem oscilações e entendem esse

número como a média dos conteúdos engarrafados. Discuta um critério para a aplicação de multas por diminuição do conteúdo engarrafado.

3. Um fabricante afirma que sua vacina previne 80% dos casos de uma certa doença. Um grupo de médicos desconfia que a vacina não é tão eficiente assim. Você poderia ajudá-los e indicar como resolver a questão?

4. Deseja-se verificar se o número de falhas de comunicação, em uma central de telefonia celular, segue o modelo Poisson com $\lambda = 10$ falhas/dia. Discuta um procedimento para decidir sobre a aceitação ou não do modelo proposto.

5. Para decidir sobre o vício de uma moeda, são feitos 100 lançamentos e o número de caras é contado. Aceita-se o equilíbrio da moeda se obtemos entre 45 e 50 caras, caso contrário, a moeda é considerada viciada. Discuta a eficiência do critério.

8.2 Teste para a Média Populacional

Vamos desenvolver as idéias gerais de teste de hipóteses supondo, inicialmente, que o modelo Normal é adequado para os dados. Situações mais gerais serão comentadas ao final desta seção.

No Exemplo 8.1 o interesse consiste em testar se a média populacional μ é igual a 14, caso em que os indivíduos pertencem à população de sadios, contra a alternativa de ser igual a 18, valor que corresponde à população de doentes. Como estamos tratando com a média populacional, utilizaremos, no teste a média amostral $\overline{X}$, um estimador não viciado e consistente de μ. Será baseado no valor observado de $\overline{X}$, denotado por $\overline{x}_{obs}$, que tomaremos nossa decisão a respeito da eficácia do tratamento proposto.

Pelas suposições feitas no Exemplo 8.1, a concentração da substância segue um modelo Normal com desvio padrão de 6 unidades/ml. Então, para o tamanho de amostra igual a 30, a média amostral terá distribuição $N(\mu, 36/30)$. Por ser uma variável aleatória, $\overline{X}$ poderá apresentar valores maiores que 14, mesmo quando $\mu = 14$. De fato, sabemos que $P(\overline{X} > 14 \mid \mu = 14) = 0{,}5$ pela simetria da distribuição Normal. Um critério que pode ser utilizado, para decidir sobre o valor de μ, é determinar um *valor crítico*, digamos x_c, tal que, se $\overline{X}$ for maior que x_c, concluímos que a amostra pertence à população com média $\mu = 18$, ou seja, o tratamento não é eficaz. Por outro lado, quando a média amostral for menor ou igual ao valor x_c, concluímos que a amostra pertence à população com média $\mu = 14$, sendo o tratamento considerado eficaz. Note que, como $\overline{X}$ é uma variável aleatória contínua, em termos probabilísticos, podemos incluir a

igualdade a x_c em qualquer uma das decisões. A figura, a seguir, resume as conclusões discutidas acima:

É importante ter em mente que, para a argumentação anterior ficar completa, precisamos determinar o valor x_c e quantificar os erros associados às possíveis conclusões. Observe que, sendo $\overline{X}$ uma variável aleatória, corremos o risco de concluir incorretamente que o tratamento é eficaz. Ou, de modo recíproco, decidir que o tratamento não é eficiente quando ele é. Portanto, é necessário estudar e quantificar os possíveis erros associados à decisão tomada.

As duas hipóteses sobre a eficácia do tratamento são denotadas por H_o e H_a e, usualmente, denominadas *hipótese nula* e *hipótese alternativa*, respectivamente. Assim,

H_o : O tratamento não é eficaz;
H_a : O tratamento é eficaz.

Essas hipóteses correspondem aos diferentes valores do parâmetro μ e, assim, podemos reescrevê-las como:

$$H_o : \mu = 18 \textit{ versus } H_a : \mu = 14.$$

As hipóteses definidas da forma acima, sem conter desigualdades, são denominadas *hipóteses simples*, mas é mais comum o uso de *hipóteses compostas*. Estas últimas podem ainda ser classificadas como unilaterais ou bilaterais, dependendo do interesse do estudo.

No caso do tratamento ser eficaz, é razoável assumirmos que ele foi capaz de fazer com que os indivíduos da amostra mudassem para uma população cuja média é inferior a 18 unidades/ml, caso contrário, se o tratamento é ineficaz, μ não se alteraria. Assim, as hipóteses de interesse seriam escritas como

$$H_o : \mu = 18 \textit{ versus } H_a : \mu < 18,$$

Nesta situação, temos o *teste de hipóteses unilateral*.

Para verificarmos se o tratamento produz algum efeito, seja ele benéfico ($\mu < 18$) ou danoso ($\mu > 18$), devemos construir um *teste de hipóteses bilateral*:

$$H_o : \mu = 18 \ \textit{versus} \ H_a : \mu \neq 18.$$

Por conveniência técnica, sempre, deixamos a igualdade na hipótese nula.

Os dois erros que podem ser cometidos ao se realizar um teste de hipóteses são:

(i) Rejeitar a hipótese H_o, quando tal hipótese é verdadeira, e

(ii) Não rejeitar a hipótese H_o quando ela deveria ser rejeitada.

Note que nenhum erro é cometido e a conclusão é correta, quando rejeitamos H_o e ela é falsa, ou, quando decidimos aceitá-la, no caso dela ser verdadeira. Ao erro cometido em (i) denominamos *erro do tipo I*, enquanto que ao erro em (ii) denominamos *erro do tipo II*. A figura, a seguir, resume as possíveis situações:

		Situação	
		H_o Verdadeira	*H_o Falsa*
Decisão	*Rejeitar H_o*	Erro Tipo I	Sem erro
	Não rejeitar H_o	Sem erro	Erro Tipo II

Figura 8.1: Erros associados a testes de hipóteses.

Como veremos adiante, uma parte importante do teste de hipóteses é controlar a probabilidade de cometermos o erro do tipo I. Essa probabilidade é denotada por α, sendo β a probabilidade de erro do tipo II. Isto é,

$$\alpha = P(\text{erro tipo I}) = P(\text{rejeitar } H_o \mid H_o \text{ verdadeira});$$
$$\beta = P(\text{erro tipo II}) = P(\text{não rejeitar } H_o \mid H_o \text{ falsa}).$$

Considerando as hipóteses $H_o : \mu = 18$ *versus* $H_a : \mu < 18$, temos a seguinte interpretação para os erros:

$$\alpha = P(\text{concluir que o tratamento é eficaz quando na verdade ele não é});$$
$$\beta = P(\text{concluir que o tratamento não é eficaz quando na verdade ele é}).$$

A situação ideal é aquela em que ambas as probabilidades, α e β, são próximas de zero. Entretanto, é fácil ver que à medida que diminuímos α, a probabilidade de erro tipo II tende a aumentar. Identifique, na Figura 8.2, as áreas relativas a α e β e veja como, dependendo do posicionamento de x_c, a diminuição da área correspondente a α implica em um aumento da área correspondente a β. Levando isso em conta, devemos cuidar para que, ao definir as hipóteses, o erro mais importante a ser evitado seja o erro do tipo I. À sua probabilidade α damos o nome de *nível de significância* do teste.

Figura 8.2: Representação gráfica dos erros α e β.

Supondo α conhecido, vamos descrever como determinar o valor crítico x_c. Inicialmente, note que

$$\begin{aligned}\alpha &= P(\text{erro tipo I}) = P(\text{rejeitar } H_o \mid H_o \text{ verdadeira}) \\ &= P(\overline{X} < x_c \mid \mu = 18) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{x_c - 18}{6/\sqrt{30}}\right) \\ &= P(Z < z_c),\end{aligned}$$

com $Z \sim N(0,1)$. Portanto, dado α obtemos z_c na tabela da Normal e calculamos x_c da seguinte forma:

$$z_c = \frac{x_c - 18}{6/\sqrt{30}} \Rightarrow x_c = 18 + z_c \frac{6}{\sqrt{30}}.$$

Por exemplo, para $\alpha = 0{,}05$ temos,

$$0{,}05 = P(Z < z_c) \Rightarrow z_c = -1{,}64;$$

logo,

$$x_c = 18 - 1{,}64 \frac{6}{\sqrt{30}} = 16{,}20.$$

Uma vez colhida a amostra, se a estimativa $\overline{x}_{obs}$ é tal que $\overline{x}_{obs} < 16{,}20$; rejeitamos a hipótese nula concluindo que o tratamento é eficaz. A região dada pelo conjunto dos números reais menores que 16,20 é denominada de *Região de Rejeição* ou *Região Crítica* e representada por *RC*. Assim,

$$RC = \{x \in \mathbb{R} : x < 16{,}20\}.$$

Denominamos *Região de Aceitação* (*RA*) ao complementar de *RC*.

Se a amostra obtida forneceu a estimativa $\overline{x}_{obs} = 16{,}04$; que pertence à *RC*, rejeitamos H_o ao nível de significância $\alpha = 0{,}05$. Graficamente a situação pode ser visualizada na Figura 8.3, em que as curvas representam a distribuição amostral de $\overline{X}$ sob a condição definida por H_o ($\mu = 18$ para a curva *Doente*) e por H_a ($\mu = 14$ para a curva *Sadio*).

Figura 8.3: Representação gráfica da região de rejeição- unilateral.

A construção de testes de hipóteses bilaterais é feita de maneira similar à apresentada para o caso unilateral, exceto que, agora, devemos considerar uma Região de Rejeição composta de duas partes disjuntas. Para exemplificar, suponha que μ_o é uma constante conhecida e que as hipóteses nula e alternativa são expressas como

$$H_o : \mu = \mu_o;$$
$$H_a : \mu \neq \mu_o.$$

A Região Crítica será dada por

$$RC = \{x \in \mathbb{R} : x < x_{c_1} \text{ ou } x > x_{c_2}\},$$

e, para um valor α fixado, determinamos os números x_{c_1} e x_{c_2} de modo que

$$P(\overline{X} < x_{c_1} \text{ ou } \overline{X} > x_{c_2}) = \alpha.$$

Dada a simetria da densidade Normal, distribuímos a massa α igualmente entre as duas partes da Região de Rejeição. Isto é,

$$P(\overline{X} < x_{c_1}) = \frac{\alpha}{2} \text{ e } P(\overline{X} > x_{c_2}) = \frac{\alpha}{2}.$$

A figura a seguir ilustra, graficamente, a escolha dos valores críticos.

Figura 8.4: Representação gráfica da região de rejeição- bilateral.

No próximo exemplo, faremos um teste de hipóteses bilateral e calcularemos a probabilidade do erro tipo II.

Exemplo 8.3: Um pequisador deseja estudar o efeito de certa substância no tempo de reação de seres vivos a um certo tipo de estímulo. Um experimento é desenvolvido com cobaias que são inoculadas com a substância e submetidas a um estímulo elétrico, com seus tempos de reação (em segundos) anotados. Os seguintes valores foram obtidos: 9,1; 9,3; 7,2; 7,5; 13,3; 10,9; 7,2; 9,9; 8,0; 8,6. Admite-se que o tempo de reação segue, em geral, o modelo Normal com média 8 e desvio padrão $\sigma = 2$ segundos. O pesquisador desconfia, entretanto, que o tempo médio sofre alteração por influência da substância. Neste caso, as hipóteses de interesse são:

H_o : *as cobaias apresentam tempo de reação padrão;*
H_a : *as cobaias têm o tempo de reação alterado.*

Em termos estatísticos, tais hipóteses envolvem o parâmetro μ e podem ser escritas como

$$H_o : \mu = 8{,}0;$$
$$H_a : \mu \neq 8{,}0.$$

Uma vez que o teste envolve a média populacional, consideramos a média amostral $\overline{X}$ para construir a estatística de teste e usamos que $\overline{X} \sim N(\mu, \frac{4}{10})$. Tendo em vista a especificação de H_a, a região crítica será da forma

$$RC = \{x \in \mathbb{R} : x < x_{c_1} \; ou \; x > x_{c_2}\}.$$

Logo, fixando $\alpha = 0{,}06$ temos:

$$\begin{aligned}
0{,}06 &= P(\text{erro tipo I}) \\
&= P(\text{rejeitar } H_o \,|\, H_o \text{ verdadeira}) \\
&= P(\overline{X} \in RC \,|\, \mu = 8{,}0) \\
&= P(\overline{X} < x_{c_1} \; ou \; \overline{X} > x_{c_2} \,|\, \mu = 8{,}0) \\
&= P\Big(\frac{\overline{X} - 8{,}0}{\sqrt{4/10}} < \frac{x_{c_1} - 8{,}0}{\sqrt{4/10}} \; ou \; \frac{\overline{X} - 8{,}0}{\sqrt{4/10}} > \frac{x_{c_2} - 8{,}0}{\sqrt{4/10}}\Big) \\
&= P(Z < z_{c_1} \; ou \; Z > z_{c_2}),
\end{aligned}$$

onde $z_{c_j} = (x_{c_j} - 8{,}0)/\sqrt{4/10}$, $j = 1, 2$ e $Z \sim N(0, 1)$. Da tabela da distribuição Normal, segue que $z_{c_1} = -1{,}88$ e $z_{c_2} = 1{,}88$. Logo,

$$x_{c_1} = 8 - 1{,}88\sqrt{4/10} = 6{,}8;$$
$$x_{c_2} = 8 + 1{,}88\sqrt{4/10} = 9{,}2.$$

Assim, podemos expressar a Região Crítica para $\alpha = 0{,}06$ como

$$RC = \{x \in \mathbb{R} : x < 6{,}8 \; ou \; x > 9{,}2\}.$$

Calculando a média amostral obtemos $\overline{x}_{obs} = 9{,}1$. Como este valor não pertence à RC, aceitamos a hipótese H_o ao nível de significância 6%. Em outras palavras, concluímos que o tempo de reação das cobaias submetidas à substância não fica alterado.

Podemos também calcular a probabilidade β do erro tipo II, isto é, a probabilidade de estar aceitando incorretamente H_o. Note que para calcular α, μ está bem especificado, o que não é o caso para o erro tipo II. Como a hipótese alternativa é composta, existem diversos valores possíveis para μ. Dessa forma, β será função de qual valor de μ foi escolhido dentro da região definida pela hipótese H_a. Nesse caso, a probabilidade do erro tipo II será denotada por $\beta(\mu)$. Por exemplo, para $\mu = 9{,}0$ teríamos

$$\begin{aligned}
\beta(9{,}0) &= P(\text{erro tipo II}) \\
&= P(\text{não rejeitar } H_o \,|\, H_o \text{ falsa}) \\
&= P(\overline{X} \notin \text{RC} \,|\, \mu = 9{,}0) \\
&= P(6{,}8 \leq \overline{X} \leq 9{,}2 \,|\, \mu = 9{,}0) \\
&= P\Big(\frac{6{,}8 - 9{,}0}{\sqrt{4/10}} \leq \frac{\overline{X} - 9{,}0}{\sqrt{4/10}} \leq \frac{9{,}2 - 9{,}0}{\sqrt{4/10}}\Big) \\
&= P(-3{,}48 \leq Z \leq 0{,}32) \\
&= 0{,}4997 + 0{,}1255 \\
&= 0{,}6252.
\end{aligned}$$

Assim, em sendo $\mu = 9{,}0$ e com probabilidade 0,6252 estaríamos concluindo, de forma equivocada, que H_o é verdadeira. □

Para caracterizar o desempenho do teste, definimos a *função poder* por $\pi(\mu) = P(\text{rejeitar } H_o \,|\, \mu)$. Note que, se o valor de μ for aquele de H_o, $\pi(\mu)$ é igual ao nível de significância α. Sendo μ um dos valores de H_a, a função poder será dada por $\pi(\mu) = 1 - \beta(\mu)$. Para um mesmo nível de significância α, quanto maior o poder melhor o teste, mas essa discussão não será feita neste texto e

recomendamos a leitura das referências mencionadas na bibliografia. Para o Exemplo 8.3, a função poder é apresentada na figura a seguir, na qual pode-se observar que à medida que nos afastamos da região da hipótese nula, o poder aumenta, atingindo valores próximos a 1. A curva apresentada é típica de um teste de hipóteses bilateral.

Figura 8.5: Função Poder.

Nos exemplos anteriores, fixamos a probabilidade do erro tipo I para encontrar a região crítica que foi utilizada como regra de decisão. Já vimos que não é possível diminuir os dois erros, simultaneamente, num mesmo teste. Por exemplo, na comparação dos grupos doente e sadio, observamos, através da Figura 8.2 que, ao diminuirmos um dos erros, estamos necessariamente aumentando o outro. Como alternativa para diminuir simultaneamente os valores de α e β, temos que recorrer a um aumento no valor de n, o tamanho da amostra. De fato, quanto maior n, melhor é a precisão do estimador utilizado e maior é o poder do teste. Observe que a variância do estimador $\overline{X}$ é σ^2/n, que diminui à medida que n cresce.

Nos testes desenvolvidos nesta seção, duas suposições básicas foram feitas: a variável aleatória de interesse na população segue o modelo Normal e sua variância é conhecida. A ausência de normalidade pode ser contornada com o

auxílio do Teorema Central do Limite o qual garante que, para amostras grandes, a média amostral tem distribuição Normal. Nesses casos, praticamente não haverá alteração nos procedimentos que estudamos até agora, pois continuamos a usar a densidade Normal para estabelecer a região crítica. Entretanto, se a variância for desconhecida, ela precisará ser estimada e precisaremos de uma nova distribuição para $\overline{X}$. Discutiremos esse caso na Seção 8.3.

Apresentamos, a seguir, um sumário das etapas para a realização de um teste de hipóteses:

1. Estabelecer as hipóteses nula e alternativa
2. Definir a forma da região crítica, com base na hipótese alternativa
3. Identificar à distribuição do estimador e obter sua estimativa
4. Fixar α e obter a região crítica
5. Concluir o teste com base na estimativa e na região crítica

Figura 8.6: Etapas de um teste de hipóteses.

Concluímos essa seção ilustrando as etapas acima descritas com um teste de proporção. As definições, técnicas e conceitos utilizadas na realização de testes de hipóteses para a média populacional são aplicáveis, quando o interesse reside na proporção de alguma característica na população. O próximo exemplo ilustra esse caso.

Exemplo 8.4: Um relatório de uma companhia afirma que 40% de toda a água obtida, através de poços artesianos no nordeste, é salobra. Há muitas controvérsias sobre essa informação, alguns dizem que a proporção é maior, outros que é menor. Para dirimir as dúvidas, 400 poços foram sorteados e observou-se, em 120 deles, água salobra. Qual seria a conclusão, ao nível de 3%?

Iniciamos com a definição das hipóteses. O parâmetro de interesse é a proporção de poços com água salobra dentre todos os poços no nordeste. Representemos essa proporção por p. Pela informação fornecida, devemos realizar um teste bilateral, com

$$H_o : p = 0{,}40\,;$$

$$H_a : p \neq 0{,}40.$$

Sabemos que o melhor estimador para p é a proporção amostral $\widehat{p}$ cuja distribuição pode ser bem aproximada por um modelo Normal, isto é, admitimos

$$\widehat{p} \sim N(p, p(1-p)/n).$$

Dado que o teste é bilateral, a Região Crítica é da forma

$$RC = \{\, x \in \mathbb{R} \mid x < p_{c_1} \; ou \; x > p_{c_2} \}.$$

Para $\alpha = 0{,}03$; os valores p_{c_1} e p_{c_2} são calculados através de

$$P(\widehat{p} < p_{c_1} \mid H_o) = \frac{0{,}03}{2} \quad \text{e} \quad P(\widehat{p} > p_{c_2} \mid H_o) = \frac{0{,}03}{2}.$$

Sob a hipótese H_o, $p = 0{,}40$ e, portanto, temos $\widehat{p} \sim N(0{,}40;\ 0{,}24/400)$. Assim,

$$P(\widehat{p} < p_{c_1} | H_o) = P\left(\frac{\widehat{p} - 0{,}40}{\sqrt{0{,}24/400}} < \frac{p_{c_1} - 0{,}40}{\sqrt{0{,}24/400}}\right)$$
$$= 0{,}015.$$

Da tabela da $N(0,1)$ segue que

$$\frac{p_{c_1} - 0{,}40}{\sqrt{0{,}24/400}} = -2{,}17.$$

Assim, obtemos $p_{c_1} = 0{,}347$. De forma análoga encontramos $p_{c_2} = 0{,}453$ e a região crítica será dada por

$$RC = \{x \in \mathbb{R} \mid x < 0{,}347 \; ou \; x > 0{,}453\}.$$

A amostra forneceu $\widehat{p}_{obs} = 120/400 = 0{,}300$ que pertence à região crítica. Dessa forma, concluímos que a hipótese nula deve ser rejeitada ao nível $\alpha = 3\%$, isto é, o relatório da companhia não está correto. ☒

Exercícios da Seção 8.2:

1. Uma variável aleatória tem distribuição Normal e desvio padrão igual a 12. Estamos testando se sua média é igual ou é diferente de 20 e coletamos uma amostra de 100 valores dessa variável, obtendo uma média amostral de 17,4.
a. Formule as hipóteses.
b. Obtenha a região crítica e dê a conclusão do teste para os seguintes níveis de significância: 1%, 2%, 4%, 6% e 8%.

2. Para uma variável aleatória com densidade Normal e desvio padrão 5, o teste da média $\mu = 10$ contra $\mu = 14$, teve a região crítica dada por $\{x \in \mathbb{R} : x > 12\}$ para uma amostra de tamanho 25. Determine as probabilidades dos erros tipo I e II.

3. Um estudo foi desenvolvido para avaliar o salário de empregadas domésticas na cidade de São Paulo. Foram sorteadas e entrevistadas 200 trabalhadoras. Admita que o desvio padrão dessa variável na cidade é de 0,8 salários mínimos.
a. Você conhece a distribuição do estimador $\overline{X}$? Se não, é possível fazer alguma suposição?
b. Deseja-se testar se a média é igual a 3 salários mínimos ou é menor. Formule as hipóteses adequadas.
c. Para um nível de significância de 3%, construa a região crítica.
d. Se a amostra forneceu média de 2,5 salários mínimos, qual seria a conclusão?

4. O consumo médio de gasolina num certo tipo de automóvel é de 15 km/litro, segundo informações da montadora. Uma revista especializada verificou o consumo em 25 desses veículos, escolhidos ao acaso, e constatou consumo médio de 14,3 km/litro. Admita que o consumo siga o modelo Normal com variância igual 9 (km/litro)^2.
a. Teste, ao nível de significância de 6%, a afirmação da montadora de que a média de consumo é igual a 15 km/litro, contra a alternativa de ser igual a 14 km/litro.
b. Determine a probabilidade do erro tipo II.

5. A vida média de uma amostra de 100 lâmpadas de certa marca é 1615 horas. Por similaridade com outros processos de fabricação, supomos o desvio padrão igual a 120 horas. Utilizando $\alpha = 5\%$, desejamos testar se a duração média de todas as lâmpadas dessa marca é igual ou é diferente de 1600 horas. Qual é a conclusão? Determine também a probabilidade do erro tipo II, se a média fosse 1620 horas.

6. Um criador tem constatado uma proporção de 10% do rebanho com verminose. O veterinário alterou a dieta dos animais e acredita que a doença diminuiu de intensidade. Um exame em 100 cabeças do rebanho, escolhidas ao acaso, indicou 8 delas com verminose. Ao nível de 8%, há indícios de que a proporção diminuiu?

8.3 Teste para a Média com Variância Desconhecida

Os testes de hipóteses e intervalos de confiança para média, que apresentamos até agora pressupõem que o valor da variância populacional é conhecido. Apesar de ser um caso particular, existem várias situações em que essa suposição é bastante razoável. Por exemplo, num processo industrial, se pudermos assegurar que uma certa máquina fornece medidas com precisão constante, teremos sua variabilidade conhecida. Uma outra situação seria aquela em que podemos utilizar resultados encontrados em outros trabalhos, ou até mesmo em experimentos estatísticos anteriores que mantenham alguma similaridade com o problema de interesse. Entretanto, no caso mais geral, quando não temos informação sobre a variância da variável aleatória que está sendo estudada, precisamos contornar essa dificuldade. Inicialmente, manteremos a suposição de que a variável aleatória de interesse tem distribuição Normal. O caso não Normal será comentado no final da seção.

Se o desvio-padrão é desconhecido, ele precisa ser estimado. Supondo que nossa amostra aleatória seja representada pelo vetor de variáveis aleatórias $(X_1, ..., X_n)$, todas elas com densidade Normal de média μ e variância σ^2. Vamos utilizar o "melhor" estimador que conhecemos para σ^2 que, como visto no Capítulo 7, é a variância amostral $S^2 = (\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)/(n-1)$.

Definindo agora a variável padronizada

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

vemos que T também é uma variável aleatória. Entretanto, apesar de $\overline{X}$ ter distribuição Normal, o denominador envolve a variável aleatória S^2, que fará com que a função densidade de T seja diferente da Normal. Esta nova densidade, que pode ser deduzida teoricamente, é denominada *t de Student* e seu parâmetro tem o nome de *graus de liberdade*, neste caso correspondendo ao total de dados menos 1. A notação utilizada será $t_{(n-1)}$ e, devido à complexidade da sua função densidade, as probabilidades são obtidas de tabelas construídas numericamente. A exemplo da Normal, o modelo *t-Student* tem densidade em forma de sino, entretanto as caudas tem maior massa que a $N(0,1)$ (veja a Figura 8.7).

Vale notar que, se o tamanho da amostra aumenta, a densidade *t-Student* converge para a Normal padrão. Por esta razão, as tabelas construídas se limitam a valores de graus de liberdade menores ou iguais a 120. Para os graus superiores a 120, as probabilidades são obtidas da tabela da distribuição Normal e

representados por "∞" na tabela do Apêndice A. Tal fato é conseqüência da consistência do estimador S^2 para σ^2, que faz com que a quantidade T se aproxime de Z à medida que aumenta o tamanho da amostra.

Figura 8.7: Densidade t- Student.

Diferentemente do teste de hipóteses, construído para o caso em que a variância é conhecida, a região crítica envolverá agora o termo S^2, que é uma quantidade aleatória. Dessa forma, amostras diferentes podem fornecer regiões críticas distintas uma vez que, possivelmente, elas produzirão estimativas diferentes para σ^2. Assim, quando a variância for desconhecida, optaremos por utilizar na região crítica valores da quantidade padronizada T. Apresentamos esse procedimento no próximo exemplo.

Exemplo 8.5: Deseja-se investigar se uma certa moléstia que ataca o rim altera o consumo de oxigênio desse órgão. Para indivíduos sadios, admite-se que esse consumo tem distribuição Normal com média 12 cm^3/min. Os valores medidos em cinco pacientes com a moléstia foram: 14,4; 12,9; 15,0; 13,7 e 13,5. Qual seria a conclusão, ao nível de 1% de significância?

O teste de interesse é:

H_o : A moléstia não altera a média de consumo renal de oxigênio;

H_a : Indivíduos portadores da moléstia têm média alterada.

Em termos da média populacional, estamos testando as hipóteses:

$$H_o : \mu = 12 \quad versus \quad H_a : \mu \neq 12,$$

e a região crítica é da forma

$$RC = \{t \in \mathbb{R} \,|\, t < t_1 \; ou \; t > t_2\}.$$

Sendo σ^2 desconhecido, usaremos o estimador $S^2 = (\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)/(n-1)$ e a quantidade t discutida anteriormente. Sendo H_o verdadeira, temos

$$T = \frac{\overline{X} - 12}{S/\sqrt{5}} \sim t_{(4)}.$$

Logo,

$$P(T < t_1) = 0,01/2 \Rightarrow t_1 = -4,604;$$
$$P(T > t_2) = 0,005 \Rightarrow t_2 = 4,604;$$

sendo o valor 4,604 obtido da tabela da distribuição *t-Student*, com 4 graus de liberdade. Assim, a região crítica será dada por

$$RC = \{t \in \mathbb{R} \,|\, t < -4,604 \; ou \; t > 4,604\}.$$

Sendo $\overline{x}_{obs} = 13,90$ e $s^2_{obs} = 0,67$; calculamos o valor padronizado

$$t_{obs} = \frac{\overline{x}_{obs} - 12}{s_{obs}/\sqrt{5}} = \frac{13,90 - 12}{0,82/\sqrt{5}} = 5,18.$$

Portanto, como $t_{obs} \in RC$, decidimos pela rejeição da hipótese nula, ou seja, a moléstia tem influência no consumo renal médio de oxigênio ao nível de 1%. □

Intervalo de confiança para μ com variância desconhecida

Quando a variância é desconhecida, construímos intervalos de confiança para a média populacional utilizando o modelo *t-Student*. O procedimento para a obtenção do intervalo é semelhante ao desenvolvido no capítulo anterior. Supondo uma amostra aleatória $X_1, \ldots, X_n$ obtida de uma população com distribuição Normal com média e variância desconhecidas, temos que

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}.$$

Desta forma, fixando-se o coeficiente de confiança γ $(0 < \gamma < 1)$ e utilizando a tabela da distribuição *t-Student* com $n-1$ graus de liberdade (ver Apêndice A), podemos obter o valor $t_{\gamma/2}$ tal que

$$P(-t_{\gamma/2} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\gamma/2}) = \gamma.$$

Logo, o intervalo com coeficiente de confiança γ para μ, com variância desconhecida, será dado por

$$\text{IC}(\mu, \gamma) = [\overline{X} - t_{\gamma/2}\frac{S}{\sqrt{n}}; \overline{X} + t_{\gamma/2}\frac{S}{\sqrt{n}}]$$

Exemplo 8.6: Considerando o exemplo anterior, uma vez que decidimos pela rejeição da hipótese nula, é boa prática fornecer um intervalo de confiança para a média populacional. Naquele exemplo foram obtidos $\overline{x}_{obs} = 13,90$ e $s^2_{obs} = 0,67$. Com $\gamma = 0,90$ obtemos, da tabela da distribuição *t-Student* com 4 graus de liberdade, $t_{\gamma/2} = 2,132$. Logo,

$$\begin{aligned}\text{IC}(\mu, 90\%) &= [13,90 - 2,132\sqrt{0,67/5}; 13,90 + 2,132\sqrt{0,67/5}\,] \\ &= [13,09; 14,71].\end{aligned}$$

Note que o intervalo de confiança encontrado não inclui o valor 12 para μ, que foi a hipótese nula no Exemplo 8.5. Dessa forma, confirma-se a conclusão de rejeição da hipótese H_o. □

Se a variável de interesse, além de ter variância desconhecida, não tiver densidade Normal, é necessário utilizar técnicas não-paramétricas para a realização do teste de média. Não apresentaremos essa metodologia aqui, entretanto, um caminho para contornar essa dificuldade é, novamente, considerar um tamanho de amostra suficientemente grande. Neste caso, é sabido que S^2 se aproxima de σ^2 de tal forma que o seu uso, juntamente com uma aplicação do Teorema Central do Limite, permite considerar $\overline{X}$ como tendo distribuição Normal, resultando em aproximações bastante satisfatórias do ponto de vista prático.

Os procedimentos adotados nos testes realizados até agora consistiram em fixar o nível de significância e, a partir dele, obter a região crítica para tomar a decisão de aceitar ou rejeitar a hipótese nula. Esse caminho é conhecido como *procedimento clássico* de testes de hipóteses. Uma alternativa, usualmente escolhida por pesquisadores e usuários da Estatística, consiste em não fixar α à priori. Este será o assunto da próxima seção.

Exercícios da Seção 8.3:

1. Com o auxílio da tabela *t-Student* calcule (se necessário, aproxime):
a. $P(-3{,}365 \leq t_5 \leq 3{,}365)$.
b. $P(|t_8| < 1{,}4)$.
c. $P(-1{,}1 \leq t_{14} < 2{,}15)$.
d. O valor de a tal que $P(t_9 > a) = 0{,}02$.
e. O valor de b tal que $P(t_{16} \leq b) = 0{,}05$.
f. O valor de c tal que $P(|t_{11}| \leq c) = 0{,}10$.
g. O valor de d tal que $P(|t_{21}| > d) = 0{,}05$.

2. Uma amostra com 10 observações de uma variável aleatória Normal forneceu média de 5,5 e variância amostral 4. Deseja-se testar, ao nível de significância de 5%, se a média na população é igual ou é menor que 6. Qual é a conclusão?

3. Admitindo que a pressão sangüínea arterial em homens siga o modelo Normal, 7 pacientes foram sorteados e tiveram sua pressão medida com os seguintes resultados: 84, 81, 77, 85, 69, 80 e 79.
a. Teste que a média é 82 contra a alternativa de ser 80. Use $\alpha = 2\%$.
b. Determine o intervalo de confiança para μ com coeficiente de confiança $\gamma = 98\%$.

4. O tempo de permanência de engenheiros recém formados no 1º emprego, em anos, foi estudado considerando um modelo Normal com média e variância desconhecidas. Por analogia com outras categorias profissionais, deseja-se testar se a média é 2 anos contra a alternativa de ser 3 anos. Para uma amostra de 15 engenheiros, a média obtida foi de 2,7 anos e o desvio padrão amostral 1,4 anos. Ao nível de 1%, qual a conclusão do teste?

5. Uma amostra de 20 observações de uma variável com distribuição Normal foi colhida, obtendo-se desvio padrão 1,1. No teste $\mu = 5$ contra $\mu > 5$, foi estabelecida a região critica $\{t \in \mathbb{R} \,|\, t > 2{,}033\}$. Determine a probabilidade do erro tipo I.

6. O número de pontos em um exame de inglês tem sido historicamente ao redor de 80. Sorteamos 10 estudantes que fizeram recentemente esse exame e observamos as notas: 65, 74, 78, 86, 59, 84, 75, 72, 81 e 83. Especialistas desconfiam que a média diminuiu e desejam testar essa afirmação através de um teste de hipóteses, com nível de significância de 5%. Fazendo as suposições necessárias, qual seria a conclusão do teste?

8.4 Nível Descritivo

Ao realizarmos um teste de hipóteses, partimos de um dado valor de α, pré-fixado, para construir a regra de decisão. Uma alternativa é deixar a cargo de quem vai utilizar as conclusões do teste a escolha do valor para a probabilidade α, que não precisará ser fixado à priori. A idéia consiste em calcular, supondo que a hipótese nula seja verdadeira, a probabilidade de se obter estimativas mais desfavoráveis ou extremas (à luz da hipótese alternativa) do que a que está sendo fornecida pela amostra. Esta probabilidade será o *nível descritivo*, denotado por α^* (ou P-valor). Valores pequenos de α^* evidenciam que a hipótese nula é falsa pois, sendo a amostra nossa ferramenta de inferência sobre a população, ela fornece uma estimativa que teria probabilidade muito pequena de acontecer, se H_o fosse verdadeira. O conceito do que é "pequeno" fica a cargo do usuário, que assim decide qual α usar para comparar com o valor obtido α^*.

Inicialmente, vamos considerar o caso do teste de hipóteses unilateral. Para H_o: $\mu = \mu_o$, a expressão de α^* depende da hipótese alternativa, isto é,

$$\alpha^* = P(\overline{X} < \overline{x}_{obs} | H_o \text{ verd.}) \text{ para } H_a : \mu < \mu_o$$

ou

$$\alpha^* = P(\overline{X} > \overline{x}_{obs} | H_o \text{ verd.}) \text{ para } H_a : \mu > \mu_o.$$

Exemplo 8.7: Uma associação de defesa do consumidor desconfia que embalagens de 450 gramas de um certo tipo de biscoito estão abaixo do peso. Para verificar tal afirmação, foram coletados ao acaso 80 pacotes em vários supermercados, obtendo-se uma média de peso de 447 gramas. Admitindo-se que o peso dos pacotes segue o modelo Normal com desvio padrão de 10 gramas, que conclusão pode ser tirada através do nível descritivo?

O teste que está sendo executado é

$H_o : \mu = 450$ (peso médio conforme previsto na embalagem);
$H_a : \mu < 450$ (peso médio abaixo do previsto na embalagem).

O valor observado na amostra foi $\overline{x}_{obs} = 447$ e as suposições feitas sobre a normalidade da variável peso implicam que $\overline{X} \sim N(\mu, 100/80)$. Então,

$$\begin{aligned}\alpha^* &= P(\overline{X} < \overline{x}_{obs} | H_o \text{ verd.}) = P(\overline{X} < 447 | \mu = 450) \\ &= P(Z < -2{,}68) = 0{,}0037.\end{aligned}$$

Portanto o nível descritivo é de 0,37%, indicando a probabilidade de que encontremos valores da estimativa mais desfavoráveis à hipótese nula. Note que o valor do nível descritivo se relaciona diretamente com o nível de significância. Neste exemplo, se tivéssemos fixado o nível de significância em qualquer valor igual ou superior a 0,37%, a conclusão seria pela rejeição de H_o, ao passo que valores inferiores a 0,37% conduziriam à aceitação da hipótese nula (ver Figura 8.8). □

Figura 8.8: Nível descritivo unilateral.

Para o teste de hipóteses bilateral, ao calcularmos o nível descritivo, precisamos considerar que a forma da região crítica envolve os valores de $\overline{x}_{obs}$ que se distanciam muito (para mais ou para menos) daquele previsto pela hipótese nula. Dessa forma, um procedimento usual é multiplicar por dois a probabilidade obtida em uma das caudas, de modo a preservar a idéia de afastamento bilateral. Assim, ao testarmos H_o: $\mu = \mu_o$ contra H_a: $\mu \neq \mu_o$, a definição do nível descritivo depende da relação entre $\overline{x}_{obs}$ e μ_o.

1º caso: se $\overline{x}_{obs} < \mu_o$, $\alpha^* = 2 \times P(\overline{X} < \overline{x}_{obs} | H_o \text{ verd.})$;

2º caso: se $\overline{x}_{obs} > \mu_o$, $\alpha^* = 2 \times P(\overline{X} > \overline{x}_{obs} | H_o \text{ verd.})$.

No caso bilateral, o nível descritivo leva em conta a posição relativa entre a estimativa $\overline{x}_{obs}$ e o valor μ_o. Desta maneira, garantimos a inclusão de valores mais extremos do estimador em relação à μ_o. A Figura 8.9 ilustra esses casos.

Figura 8.9: Regiões desfavoráveis num teste bilateral com H_o: $\mu = \mu_o$.

Exemplo 8.8: Vamos considerar o teste apresentado no Exemplo 8.3. As hipóteses sobre o tempo de reação de cobaias, submetidas a um estímulo elétrico, foram as seguintes:

$H_o : \mu = 8$ (tempo médio de reação sem alteração) ;
$H_a : \mu \neq 8$ (tempo médio de reação alterado).

Para uma amostra de 10 cobaias, observou-se $\overline{x}_{obs} = 9{,}1$. Com as suposições já feitas naquele exemplo (normalidade com $\sigma = 2$ segundos), podemos obter o nível descritivo. Note que aqui os valores da estimativa mais desfavoráveis em relação à H_o correspondem a região $\overline{X} > 9{,}1$. Assim, temos

$$\begin{aligned}\alpha^* &= 2 \times P(\overline{X} > \overline{x}_{obs} | H_o \text{ verd.}) \\ &= 2 \times P(\overline{X} > 9{,}1 | \mu = 8) \\ &= 2 \times P(Z > 1{,}74) \\ &= 0{,}0818.\end{aligned}$$

Logo, se desejarmos utilizar um nível de significância igual a 0,05 concluiríamos pela aceitação da hipótese H_o, ao passo que um nível de significância igual a 0,10 nos levaria a rejeitar a hipótese H_o (ver Figura 8.10).

Figura 8.10: Nível descritivo bilateral.

O nível descritivo nos fornece uma idéia da intensidade com a qual estamos rejeitando, ou não, a hipótese nula. Dessa forma, tem papel importante do ponto de vista exploratório, uma vez que pode nos fornecer indicações para pesquisas futuras. □

Exercícios da Seção 8.4:

1. Um pesquisador está realizando um teste para a média e obteve nível descritivo igual a 0,035. Ele aceitará a hipótese nula para níveis de significância superiores ou inferiores à 0,035?

2. Uma variável aleatória tem distribuição Normal e desvio padrão igual a 10. Uma amostra de 50 valores dessa variável forneceu média igual a 15,2. Para cada um dos testes abaixo responda qual é o nível descritivo.
   a. $H_o : \mu = 18$ *versus* $H_a : \mu = 13$.
   b. $H_o : \mu = 18$ *versus* $H_a : \mu < 18$.
   c. $H_o : \mu = 18$ *versus* $H_a : \mu \neq 18$.
   d. $H_o : \mu = 17$ *versus* $H_a : \mu = 14$.

3. A resistência de um certo tipo de cabo de aço é uma variável aleatória modelada pela distribuição Normal com desvio padrão 6 kgf. Uma amostra de tamanho 25 desses cabos, escolhida ao acaso, forneceu média igual a 9,8 kgf. Para o teste $\mu = 13$ contra $\mu = 8$, qual é o nível descritivo? Que conclusão você consideraria adequada?

4. Sorteamos, ao acaso, 12 observações de uma variável aleatória que segue o modelo Normal. Da amostra obtivemos média 21,7 e desvio padrão 5,5. Determine o nível descritivo do teste $\mu = 18$ contra $\mu > 18$.

8.5 Testes Qui-Quadrado

Apresentamos, nesta seção, três testes que utilizam o modelo Qui-Quadrado como estrutura probabilística e, por essa razão, são denominados, de forma geral, *Testes Qui-Quadrado*. Iniciamos testando a adequabilidade de um modelo probabilístico para uma dada situação, depois discutimos o teste de independência entre duas variáveis e encerramos a seção com o teste de homogeneidade de subpopulações.

Nas seções anteriores, nosso problema foi testar hipóteses sobre os parâmetros média e proporção. Em geral, as formas das distribuições de probabilidade eram conhecidas (ou seriam aproximadas) e tínhamos que decidir quanto a aceitar uma ou outra hipótese, sobre o valor desse parâmetro. Em termos práticos, outra situação comum é termos observações de uma variável aleatória cuja distribuição na população é desconhecida. Nesse caso, uma das primeiras providências é tentar identificar o comportamento da variável com um modelo teórico. Em algumas situações, é possível incorporar informações de outras variáveis que descrevam fenômenos aleatórios similares e tenham distribuição conhecida. Dessa forma, teríamos um candidato a modelo e nosso problema seria estabelecer um procedimento para aceitá-lo ou não. Existem, contudo, vários outros casos em que não se tem a menor idéia do comportamento da variável. Uma das maneiras iniciais de análise é construir um diagrama, com as freqüências de ocorrência, nos moldes do histograma. Dessa representação gráfica, pode sair a sugestão de modelos adequados aos dados. Em qualquer caso, o modelo proposto pode ser testado através do chamado *Teste de Aderência*. Nesta seção, apresentaremos um desses testes que usa a distribuição Qui-Quadrado, outros testes de aderência podem ser encontrados nas referências mencionadas na bibliografia.

Considere uma variável X para a qual temos uma amostra de valores e deseja-se verificar a adequação ou não de um certo modelo probabilístico. Os valores observados da variável foram divididos em k categorias contendo, cada uma, um ou mais valores que são apresentados numa tabela de freqüência:

Categoria	1	2	3	$\cdots$	k
Freq. Observada	o_1	o_2	o_3	$\cdots$	o_k

Se X for discreta, as categorias são os próprios valores da variável, eventualmente agregando mais de um valor na mesma categoria. No caso contínuo, as categorias são definidas a partir de faixas de valores da variável. Do modelo que está sendo sugerido, calculamos as freqüências esperadas em cada uma das categorias. Assim,

Categoria	1	2	3	$\cdots$	k
Freq. Esperada	e_1	e_2	e_3	$\cdots$	e_k

Se X seguir o modelo proposto, essas duas tabelas não devem ser muito discrepantes. O teste de aderência cria, então, o critério, para decidir se podemos aceitar ou não o modelo indicado. Em outras palavras, decidimos se os dados amostrais *aderem* ao modelo ou não. As hipóteses do teste são:

H_o: X segue o modelo proposto;
H_a: X não segue esse modelo.

A quantidade que usaremos para tomar nossa decisão será baseada na diferença entre os valores esperados sob H_o e aqueles observados na amostra. Podemos dizer que a diferença $o_i - e_i$ dá uma idéia da compatibilidade entre os valores observados e o modelo proposto. Assim, se as diferenças forem muito grandes, é razoável admitir que o modelo não deve ser adequado. Por outro lado, pequenas diferenças podem ser aceitas, pois estamos sempre sujeitos a flutuações, quando trabalhamos com variáveis aleatórias. Baseando-se nessa idéia intuitiva, a quantidade utilizada no teste será:

$$Q^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

sendo que k representa o número de categorias, o_i a freqüência observada e e_i a freqüência esperada para a categoria i.

Para interpretar a expressão de Q^2, note que o termo $o_i - e_i$ indica a diferença, na categoria i, entre a freqüência observada e a esperada ou, em outras palavras, o desvio em relação ao modelo proposto. Se, simplesmente, fizéssemos a soma desses desvios para todas as categorias, obteríamos zero, pois o total de dados é o mesmo. Para evitar isso, tomamos o quadrado dos desvios. Entretanto, por serem quantidades não negativas, sua soma poderia se tornar artificialmente alta e, por essa razão, é conveniente fazermos uma mudança de escala dividindo esses desvios ao quadrado pela freqüência esperada. Somando agora, para todas

as categorias, obtemos a expressão de Q^2 que é, assim, uma medida da discrepância que queremos quantificar.

É possível demonstrar que, para um tamanho de amostra suficientemente grande, a distribuição de Q^2 pode ser aproximada por um modelo Qui-Quadrado com parâmetro $k-1$, denominado de *número de graus de liberdade* da distribuição. Essa distribuição é representada por $\mathcal{X}^2_{(k-1)}$. O modelo Qui-Quadrado é contínuo e assume valores não negativos. Sua densidade tem uma expressão complexa de forma que probabilidades serão obtidas da tabela apresentada no Apêndice. A aproximação para o modelo Qui-Quadrado será melhor, se todas as freqüências esperadas forem ao menos iguais a 5. Se isto não acontecer para alguma categoria, devemos combiná-la a uma outra de forma conveniente, garantindo que todas as freqüências esperadas atendam a esse critério. Retomamos agora o Exemplo 8.2, construindo formalmente o teste de aderência.

Exemplo 8.9: No Exemplo 8.2, definimos X como sendo o *número de impactos anteriores à falha* em um equipamento eletrônico. Uma amostra de 80 ensaios foi obtida, cada ensaio representando os testes feitos até a interrupção por falha no equipamento, resultando 80 observações da variável de interesse. Pretende-se verificar se o modelo Geométrico com $p = 0{,}4$ é adequado. O teste será:

H_o: $X \sim G(0{,}4)$;
H_a: X tem outra distribuição.

A decisão será baseada no comportamento de Q^2, definido acima. Considerando o tamanho de amostra grande, a distribuição de Q^2 pode ser aproximada pela Qui-Quadrado, com número de graus de liberdade que depende de quantas categorias serão estabelecidas. A região crítica é constituída de valores grandes de Q^2, isto é,

$$RC = \{w : w \geq q_c\},$$

com q_c sendo determinado pelo nível de significância do teste, ou seja,

$$\alpha = P(Q^2 \geq q_c \,|\, H_o \text{ verdadeiro}).$$

Para determinar o valor observado de Q^2, denotado por q^2_{obs}, precisamos obter as freqüências esperadas. Se H_o for verdadeiro, X segue o modelo Geométrico, isto é, $P(X = k) = p_k = 0{,}4 \times 0{,}6^k$. Logo,

$$\text{Freq. esperada de resistência a } k \text{ impactos} = 80 \times p_k = 80 \times 0{,}4 \times 0{,}6^k.$$

Na tabela, a seguir, apresentamos as freqüências esperadas e os valores que foram observados no teste de resistência realizado.

Impactos	0	1	2	3	4	mais de 4
Freq. observada	30	26	10	5	5	4
Freq. esperada	32,0	19,2	11,5	6,9	4,1	6,3

Como a categoria correspondente ao valor 4 tem freqüência esperada igual a 4,1 que é menor que 5, agregamos as duas últimas categorias formando a dos *maiores de 3*, a qual terá a freqüência observada de 9 e esperada de 10,4. Então,

$$q^2_{obs} = \frac{(30-32{,}0)^2}{32{,}0} + \frac{(26-19{,}2)^2}{19{,}2} + \cdots + \frac{(9-10{,}4)^2}{10{,}4} = 3{,}44.$$

Escolhendo $\alpha = 0{,}05$; vamos determinar q_c utilizando a distribuição Qui-Quadrado, com 4 graus de liberdade. Temos,

$$P(Q^2 \geq q_c \,|\, H_o) = \alpha \Rightarrow P(\, Q^2 \geq q_c \,|\, H_o) = 0{,}05.$$

Consultando a tabela na linha correspondente a 4 graus de liberdade e na coluna de 5%, o valor crítico será $q_c = 9{,}49$ que é maior que o valor observado de 3,44. Concluímos pela aceitação do modelo proposto. A próxima figura apresenta a densidade do modelo $\mathcal{X}^2_4$ com a região crítica do teste. □

Figura 8.11: Densidade $\mathcal{X}^2_4$ e Região Crítica.

Uma situação bastante comum é aquela em que desejamos testar se uma variável segue um certo modelo, mas desconhecemos um ou mais parâmetros da distribuição. Sendo assim, vamos utilizar a amostra para chegarmos às estimativas dos parâmetros desconhecidos, isto é, utilizando as próprias observações que dispomos, vamos obter estimativas que serão consideradas como valores dos parâmetros desconhecidos. Nesses casos, o número de graus de liberdade se altera para $k-1-e$, com e representando o número de parâmetros que foram estimados. No próximo exemplo, ilustramos essa situação ao testar a aderência de um conjunto de observações a um modelo contínuo.

Exemplo 8.10: Deseja-se verificar a afirmação de que a porcentagem de cinzas contidas em carvão, produzido por uma certa empresa, segue a distribuição Normal. Os dados, apresentados a seguir, representam a quantidade percentual de cinzas encontradas em 250 amostras de carvão analisadas em laboratório.

Cinzas (em %)	freq. observada
9,5 ⊢ 10,5	2
10,5 ⊢ 11,5	5
11,5 ⊢ 12,5	16
12,5 ⊢ 13,5	42
13,5 ⊢ 14,5	69
14,5 ⊢ 15,5	51
15,5 ⊢ 16,5	32
16,5 ⊢ 17,5	23
17,5 ⊢ 18,5	9
18,5 ⊢ 19,5	1

Qual decisão devemos tomar ao nível de significância de 4%?

Como desconhecemos a média e a variância da Normal que será testada, precisamos, inicialmente, obter suas estimativas a partir da amostra. Os melhores estimadores desses parâmetros são a média e a variância amostral, representados por $\overline{X}$ e S^2, respectivamente. Para calcularmos suas estimativas, tomamos o ponto médio do intervalo como representante dos valores da respectiva classe. Então, temos

$$\widehat{\mu} = \overline{x}_{obs} = 14{,}5;$$
$$\widehat{\sigma}^2 = s^2_{obs} = 2{,}7.$$

Denominando por C a variável aleatória: *porcentagem de cinzas contidas no*

carvão produzido pela empresa, as hipóteses a serem testadas são:

$$H_o : C \sim N\,(14{,}5; 2{,}7);$$
$$H_a : C \text{ tem outra distribuição.}$$

Como antes, usaremos a estatística Q^2 para tomar a decisão e, considerando o tamanho da amostra grande o suficiente, aproximamos a distribuição de Q^2 pela Qui-Quadrado. Dessa forma, utilizando $\alpha = 4\%$, obteremos a região crítica do teste.

As diversas faixas constituem as categorias de valores da variável C e serão numeradas de 1 a 10. De modo a varrer os valores do intervalo $(-\infty, \infty)$, correspondentes ao modelo Normal, acrescentamos às categorias 1 e 10 os valores, respectivamente, menores que 9,5 e maiores que 19,5. Dessa forma, para calcular as freqüências esperadas, procedemos da seguinte forma:

$$e_1 = 250 \times P(C < 10{,}5 \,|\, H_o \text{ verdadeiro});$$

$$e_i = 250 \times P(C \in \text{categoria } i \,|\, H_o \text{ verdadeiro}),\ i = 2, \ldots, 9;$$

$$e_{10} = 250 \times P(C \geq 18{,}5 \,|\, H_o \text{ verdadeiro}).$$

As probabilidades acima são calculadas da maneira usual através da tabela da Normal padrão. Por exemplo,

$$\begin{aligned} P(C < 10{,}5 \,|\, H_o \text{ verdadeiro}) &= P\left(\frac{C - 14{,}5}{\sqrt{2{,}7}} < \frac{10{,}5 - 14{,}5}{\sqrt{2{,}7}}\right) \\ &= P(Z < -2{,}44) \\ &= 0{,}0073. \end{aligned}$$

Para a categoria 5,

$$\begin{aligned} P(13{,}5 \leq C < 14{,}5 | H_o) &= P\left(\frac{13{,}5 - 14{,}5}{\sqrt{2{,}7}} \leq \frac{C - 14{,}5}{\sqrt{2{,}7}} < \frac{14{,}5 - 14{,}5}{\sqrt{2{,}7}}\right) \\ &= P(-0{,}61 \leq Z < 0) \\ &= 0{,}2291. \end{aligned}$$

As freqüências esperadas são apresentadas na tabela, a seguir, e devem somar 250, o que foi considerado nos arredondamentos efetuados.

Categoria	freq. esperada
1	1,82
2	6,58
3	19,40
4	39,92
5	57,28
6	57,28
7	39,92
8	19,40
9	6,58
10	1,82

Observamos que as categorias 1 e 10 devem ser anexadas a outras, pois suas freqüências esperadas são menores que 5. Desse modo, agrupamos as categorias 1 com 2 e 9 com 10. As novas categorias e suas respectivas freqüências esperadas e observadas são apresentadas na tabela a seguir.

Categoria	freq. esperada	freq. observada
1	8,40	7
2	19,40	16
3	39,92	42
4	57,28	69
5	57,28	51
6	39,92	32
7	19,40	23
8	8,40	10

Efetuando o cálculo da estatística Q^2, obtemos $q^2_{obs} = 6{,}57$. Para determinar a região crítica, utilizamos a distribuição Qui-Quadrado com $8 - 1 - 2 = 5$ graus de liberdade, em que perdemos dois graus de liberdade devido à estimação dos parâmetros μ e σ^2. Com o auxílio da tabela da Qui-Quadrado, obtemos

$$RC = \{w : w \geq 11{,}64\},$$

para $\alpha = 4\%$. Note que RC não contém q^2_{obs} e, portanto, decidimos pela aceitação do modelo Normal para a variável aleatória C. □

Apresentamos, agora, uma forma de testar a independência entre duas variáveis. Se dispomos da função de probabilidade conjunta de duas variáveis

aleatórias, podemos verificar se, para todos os possíveis valores das variáveis, o produto das probabilidades marginais é igual à probabilidade conjunta.

Na situação mais comum em que não temos informação sobre a ocorrência conjunta das variáveis aleatórias, o procedimento usual é coletar uma amostra anotando a freqüência conjunta da ocorrência dos valores das variáveis. Pode-se, então, utilizar um teste de hipóteses conhecido como *Teste de Independência*. Este teste será apresentado através do próximo exemplo.

Exemplo 8.11: A tabela abaixo contém os resultados obtidos por estudantes do ensino médio, em um exame com questões nas disciplinas de física e matemática. Deseja-se testar se existe dependência entre as notas dessas duas disciplinas que, para efeito de apresentação na tabela e análise de comportamento, foram classificadas nas categorias alta, média e baixa.

Física \ Matemática	Alta	Média	Baixa	Total
Alta	56	71	12	139
Média	47	163	38	248
Baixa	14	42	85	141
Total	117	276	135	528

Iremos testar as hipóteses:

H_o : As notas de física e matemática são independentes;
H_a : Elas não são independentes.

De modo análogo ao que fizemos no teste de aderência, vamos construir uma tabela de valores esperados. Para a casela (i, j), esse valor é:

$$e_{i,j} = \frac{\text{Total da linha } i \times \text{Total da coluna } j}{\text{Total geral}}.$$

Note que os valores esperados são calculados sob a hipótese H_o de independência e, por essa razão, utilizamos os totais de linha e coluna que representam as freqüências marginais das variáveis. Por exemplo, para a casela $(1, 2)$, temos:

$$e_{1,2} = \frac{\text{Total da linha 1} \times \text{Total da coluna 2}}{\text{Total geral}} = \frac{139 \times 276}{528} = 72{,}66\,.$$

A tabela completa de valores esperados é

Física \ Matemática	Alta	Média	Baixa
Alta	30,80	72,66	35,54
Média	54,95	129,64	63,41
Baixa	31,25	73,70	36,05

Para medir a diferença entre os valores esperados e observados, usaremos:

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(\mathrm{o}_{i,j} - \mathrm{e}_{i,j})^2}{\mathrm{e}_{i,j}}\,,$$

com r e s representando o número de linhas e de colunas, respectivamente. A argumentação para sua utilização é a mesma já apresentada no teste de aderência e, para um número grande de observações, a distribuição de Q^2 se comporta como um modelo Qui-Quadrado com $(r-1) \times (s-1)$ graus de liberdade. A região crítica contém valores grandes de Q^2, isto é,

$$RC = \{w : w \geq q_c\},$$

com q_c sendo determinado pelo nível de significância do teste, ou seja,

$$\alpha = P(Q^2 \geq q_c \mid H_o \text{ verdadeiro}).$$

Para $\alpha = 0{,}01$ a tabela da Qui-Quadrado com 4 graus de liberdade fornece $q_c = 13{,}28$. Obtemos assim,

$$RC = \{w : w \geq 13{,}28\}.$$

Vamos calcular o valor observado de Q^2,

$$q^2_{obs} = \frac{(56-30{,}80)^2}{30{,}80} + \frac{(71-72{,}66)^2}{72{,}66} + \cdots + \frac{(85-36{,}05)^2}{36{,}05} = 145{,}78\,.$$

Concluímos pela rejeição da hipótese nula, ou seja, as notas de física e matemática não são independentes. □

Na construção da tabela de valores esperados, caso alguma casela tenha valor menor que 5, será necessário agrupar categorias. Este procedimento visa garantir uma melhor aproximação para o uso do modelo Qui-Quadrado para Q^2.

Consideremos agora o chamado *Teste de Homogeneidade*. Esse teste consiste em verificar se uma variável aleatória se comporta de modo similar, ou homogêneo, em várias subpopulações. Apesar da mecânica de realização do teste ser semelhante a do Teste de Independência, uma distinção importante se refere à forma como as amostras são coletadas. No teste de homogeneidade, fixamos o tamanho da amostra em cada uma das subpopulações e, então, selecionamos uma amostra de cada uma delas. Na tabela apresentada a seguir, as linhas representam as subpopulações e, as colunas, os diferentes valores ou categorias da variável.

Subpopulações	Valores da variável			total de linha
1	o_{11}	o_{12}	$\cdots$	n_1
2	o_{21}	o_{22}	$\cdots$	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
total de coluna				total Geral

Para o cálculo dos valores esperados (supondo homogeneidade entre as subpopulações), utilizamos, para a casela (i, j),

$$e_{i,j} = n_i \times \frac{\text{total da coluna } j}{\text{total geral}}.$$

O total de linha n_i indica o tamanho da amostra da subpopulação i, ao passo que o quociente, total da coluna j dividido pelo total geral, representa a proporção de ocorrências do valor da variável correspondente à coluna j. Caso haja homogeneidade de comportamento da variável, esperamos que essa proporção seja a mesma, em todas as subpopulações. No próximo exemplo, apresentamos mais detalhes.

Exemplo 8.12: Estamos interessados em saber se a preferência por certo tipo de filme se altera com o estado civil. Selecionamos pessoas em cada uma das subpopulações: solteiro, casado, divorciado e viúvo. Os resultados estão na tabela a seguir:

Estado Civil \ Filme	Policial	Comédia	Romance	tam. amostra
Solteiro	45	25	30	100
Casado	36	61	43	140
Divorciado	39	36	35	110
Viúvo	14	19	17	50
total	134	141	125	400

Na tabela anterior, a última coluna representa o tamanho da amostra selecionada em cada subpopulação. Observe que esses valores foram fixados antes da coleta ser realizada. As hipóteses a serem testadas são:

H_o : A preferência por certo tipo de filme é igual para qualquer estado civil;
H_a : A preferência muda.

A proporção dos indivíduos que preferem filmes policiais é 134/400. Se a variável Filme for homogênea entre as subpopulações de Estado Civil, devemos ter essa mesma preferência por filmes policiais, para qualquer estado civil. Logo, o valor esperado de preferência pelo gênero Policial, na subpopulação dos solteiros, deve ser $100 \times 134/400$. Para as outras subpopulações, multiplicamos $134/400$ pelos respectivos valores do tamanho de amostra, que são diferentes nesse exemplo. A tabela de freqüências esperadas é apresentada a seguir:

Estado Civil \ Filme	Policial	Comédia	Romance	tam. amostra
Solteiro	33,50	35,25	31,25	100
Casado	46,90	49,35	43,75	140
Divorciado	36,85	38,78	34,37	110
Viúvo	16,75	17,62	15,63	50
total	134	141	125	400

Calculamos a quantidade Q^2 da mesma forma como fizemos anteriormente, isto é, vamos quantificar a "distância" entre os valores observados e aqueles esperados, se houvesse homogeneidade. Assim,

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(\mathrm{o}_{i,j} - \mathrm{e}_{i,j})^2}{\mathrm{e}_{i,j}}.$$

Para um número grande de observações, a distribuição de Q^2 é Qui-Quadrado com $(r-1) \times (s-1)$ graus de liberdade (r, número de linhas e s de colunas). A região crítica contém valores grandes de Q^2, isto é,

$$RC = \{w : w \geq q_c\},$$

com q_c sendo determinado pelo nível de significância do teste, ou seja,

$$\alpha = P(Q^2 \geq q_c \mid H_o \text{ verdadeiro}).$$

Escolhendo $\alpha = 0{,}05$ obtemos, da tabela da densidade Qui-Quadrado com 6 graus de liberdade, $q_c = 12{,}59$. Portanto,

$$RC = \{w : w \geq 12{,}59\}.$$

Para o valor observado de Q^2 temos:

$$q^2_{obs} = \frac{(45-33{,}50)^2}{33{,}50} + \frac{(36-46{,}90)^2}{46{,}90} + \cdots + \frac{(17-15{,}63)^2}{15{,}63} = 13{,}29.$$

Concluímos pela rejeição da hipótese nula, ou seja, a preferência de filmes não é a mesma nas diferentes subpopulações definidas pelo estado civil. □

Exercícios da Seção 8.5:

1. Utilizando a tabela da distribuição Qui-Quadrado determine (aproxime se necessário):

a. $P(\mathcal{X}^2_7 > 14{,}70)$.

b. $P(\mathcal{X}^2_{23} > 39)$.

c. $P(\mathcal{X}^2_{12} \leq 9)$.

d. $P(12 < \mathcal{X}^2_{17} \leq 30{,}2)$.

e. O valor de a tal que $P(\mathcal{X}^2_{13} > a) = 0{,}05$.

f. O valor de b tal que $P(\mathcal{X}^2_4 > b) = 0{,}01$.

g. O valor de c tal que $P(\mathcal{X}^2_{21} < c) = 0{,}95$.

2. Um pediatra pretende avaliar se o sexo de bebês pode ser modelado por uma distribuição de Bernoulli, com $p = 0{,}55$ indicando a probabilidade de nascimento de meninas. Uma amostra aleatória de 25 nascimentos indicou 13 meninas.

a. Formule as hipóteses adequadas.

b. Qual é a conclusão ao nível de significância de 5%?

3. Quatro máquinas de grande porte trabalham de forma independente e, ao fim da jornada de trabalho, são vistoriadas pelo controle de qualidade. Caso necessitem, serão ajustadas. Das informações arquivadas pela empresa, sorteamos 22 dias e anotamos o número de máquinas que sofreram ajuste nesses dias. Os dados são apresentados na tabela abaixo. O engenheiro de manutenção pretende verificar se é adequado o modelo Binomial com $n = 4$ e probabilidade de ajuste $p = 0{,}1$. Use um nível de significância de 4%.

Ajustes diários	0	1	2	3	4
Freqüência	13	6	2	1	0

4. Para verificar a qualidade do processo de fabricação, cabos de aço são submetidos ao ensaio de tração até acontecer a ruptura. Os cabos têm 5 metros de comprimento e deseja-se testar se o modelo Uniforme Contínuo é adequado. Para 30 cabos, sorteados ao acaso, obtivemos a seguinte tabela:

Faixa	freqüência
$0 \vdash 1$	7
$1 \vdash 2$	6
$2 \vdash 3$	4
$3 \vdash 4$	6
$4 \vdash 5$	7

Qual é a decisão para uma probabilidade de erro tipo I de 0,02?

5. Em um experimento para verificar a relação entre crises de asma e incidência de gripe, 150 crianças foram escolhidas, ao acaso, dentre aquelas acompanhadas pelo Posto de Saúde do bairro. Os dados referentes a uma semana são apresentados na tabela abaixo.

Asma \ Gripe	Sim	Não
Sim	27	34
Não	42	47

Você acha que as ocorrências de asma e gripe são independentes? Use $\alpha = 4\%$.

6. A opinião sobre o atendimento de pacientes com AIDS em hospitais públicos foi estudada em duas cidades. Na cidade A, sorteou-se 150 usuários e, em B, 200. Com os resultados apresentados na tabela abaixo, você diria que a opinião é a mesma nas duas cidades? Use $\alpha = 5\%$.

Cidade \ Atendimento	Bom	Regular	Ruim
A	73	37	40
B	94	61	45

7. Em uma faculdade, o desempenho esportivo dos alunos está sendo estudado para dois cursos diferentes. Os cursos de administração e economia forneceram amostras que estão representadas abaixo. Você acredita que os alunos desses dois cursos têm o mesmo desempenho, ao nível 4%?

Cursos \ Desempenho	Bom	Regular	Ruim	Total
Administração	65	70	45	180
Economia	27	103	20	150

8.6 Exercícios

1. Suponha que queiramos testar $H_o : \mu = 50$ versus $H_a : \mu > 50$, onde μ é a média de uma variável aleatória Normal com desvio-padrão igual a 10. Extraída uma amostra de $n = 36$ elementos da população, observou-se $\overline{x}_{obs} = 53$. Faça o teste utilizando os níveis 1%, 2%, 5% e 10 %.

2. Uma fábrica de automóveis anuncia que seus carros consomem, em média, 10 litros de gasolina por 100 quilômetros, com desvio padrão de 0,8 litros. Uma revista desconfia que o consumo é maior e resolve testar essa afirmação. Para tal, analisa 35 automóveis dessa marca, obtendo como consumo médio 10,2 litros por 100 quilômetros. Considerando que o consumo siga o modelo Normal, o que a revista pode concluir sobre o anúncio da fábrica ao nível de 1%? Qual o erro tipo II se a média for 10,6?

3. Uma máquina deve produzir peças com diâmetro de 2 cm. Entretanto, variações acontecem e vamos assumir que o diâmetro dessas peças siga o modelo Normal com variância igual a 0,09 cm^2. Para testar se a máquina está bem regulada, uma amostra de 100 peças é coletada.
   a. Formule o problema como um teste de hipóteses.
   b. Qual seria a região crítica se $\alpha = 0{,}02$?
   c. Se a região de aceitação fosse $\{x \in \mathbb{R} \mid 1,95 \leq x \leq 2,05\}$, qual seria o nível de significância do teste? Nesse caso, determine a probabilidade do erro tipo II se $\mu = 1{,}95$ cm.
   d. Se para essa amostra, $\overline{x}_{obs} = 1{,}94$; qual a decisão em (b)? E em (c)?

4. O atual tempo de travessia com balsas entre Santos e Guarujá é considerado uma variável aleatória com distribuição Normal de média 10 minutos e desvio padrão 3 minutos. Uma nova balsa vai entrar em operação e desconfia-se que será mais lenta que as anteriores, isto é, haverá aumento na média especificada no modelo acima.
   a. Especifique as hipóteses em discussão.
   b. Interprete os erros tipo I e tipo II.
   c. Para uma amostra de 20 tempos de travessia com a nova balsa, obtenha a região crítica considerando um nível de 5%.
   d. Calcule a probabilidade do erro tipo II, se a nova balsa demora, em média, 2 minutos a mais que as anteriores para completar a travessia.

5. O nível de colesterol no sangue é uma variável com distribuição Normal, de média μ desconhecida e desvio-padrão $\sigma = 60$ mg/100 ml.
   a. Suponha que várias amostras de tamanho n são escolhidas ao acaso desta população. Para cada indivíduo, o nível de colesterol é obtido e a média de cada uma das amostras é calculada. Qual deve ser o valor de n para que apenas 10% das médias amostrais excedam a média populacional em 10 unidades ou mais?
   b. Teste a hipótese de que $\mu = 260$, contra a alternativa de que $\mu > 260$ com base numa amostra de 50 pacientes, em que se observou uma média amostral $\overline{x}_{obs} = 268$. Utilize um nível de 5%.
   c. Qual deve ser o tamanho da amostra, escolhida na população acima, para que o intervalo de confiança para μ tenha um comprimento de 30 unidades? Use $\gamma = 99\%$.
   d. Para o teste especificado em (b), calcule a probabilidade β para o erro de tipo II, se o valor real de μ for igual a 290.

6. Suponhamos que o tempo de cura para um doente tratado pelo método A obedeça a uma distribuição Normal, com média de 7 dias e desvio-padrão de 2 dias. Um novo tratamento B é proposto com a finalidade de diminuir o tempo de cura desse tipo de paciente. Em um experimento clínico, 25 pacientes com a doença receberam o novo tratamento B e observou-se que a média do tempo de restabelecimento para eles foi de 6 dias.
   a. Sabendo que o novo tratamento não influi na variância, identifique as hipóteses adequadas e teste-as, considerando um nível de significância $\alpha = 0{,}02$.
   b. Construa um intervalo de confiança ($\gamma = 95\%$) para a verdadeira média da distribuição do tempo de cura sob o tratamento B.

7. Uma empresa fabrica cilindros com 50 mm de diâmetro. O desvio-padrão dos diâmetros produzidos é de 3,0 mm. A fim de saber se a produção encontra-se dentro dos padrões esperados, a cada hora, 4 cilindros são amostrados e têm seus diâmetros medidos. A média dos diâmetros é usada para decidir se o processo de fabricação está operando satisfatoriamente. Assim, se o diâmetro médio estiver entre 47 e 53 mm, o processo deve continuar, caso contrário, a produção é interrompida e ajustes são feitos. Suponha que o comprimento dos diâmetros é bem modelado por uma distribuição Normal.

a. Qual é a probabilidade de se parar incorretamente a produção, se a média do diâmetro continuar em 50 mm?

b. Qual é a probabilidade da produção continuar, se a média do diâmetro se deslocar para $\mu = 52$?

8. Sabe-se que a concentração média de cloro encontrada na urina de recém-nascidos, com gestação de 9 meses, é igual a 210 unidades e que o desvio-padrão correspondente é igual a 20 unidades. Sabe-se também que, em recém-nascidos prematuros, a concentração de cloro na urina tem um desvio-padrão igual àquele observado para os outros recém-nascidos, porém suspeita-se que a concentração média seja menor. Para testar a veracidade desta suspeita, uma amostra de recém-nascidos prematuros será observada com relação às concentrações de cloro na urina (admita que siga o modelo Normal).

a. Formule as hipóteses adequadas.

b. Quantos recém-nascidos prematuros devem ser observados para que tenhamos simultaneamente $\alpha = 10\%$ e $\beta_{(200)} = 5\%$.

c. Obtenha o nível descritivo do teste, se a concentração média de cloro observada na urina de uma amostra de 25 prematuros foi de 200 unidades. Interprete.

9. Um laboratório que fabrica comprimidos analgésicos anuncia que seu remédio contra dor de cabeça leva em média 14 min para aliviar a dor, com desvio-padrão de 5 min. Um médico sustenta que o tempo é maior e seleciona aleatoriamente 40 pacientes. Pede a eles que tomem tais pílulas quando tiverem dor de cabeça, anotando o tempo (em minutos) até o alivio da dor. Após coletar todas as respostas, ele verifica que o tempo médio de alivio para esses pacientes foi de 19 min. Estes resultados confirmam a afirmação feita pelo laboratório? Faça as suposições necessárias e use $\alpha = 5\%$.

10. Considere o teste $p = 0{,}6$ contra $p \neq 0{,}6$. Sendo $n = 100$, indique a probabilidade de erro tipo I para as seguintes regiões críticas:

a. $RC = \{x \in \mathbb{R} \mid x < 0{,}56 \; ou \; x > 0{,}64\}$.

b. $RC = \{x \in \mathbb{R} \mid x < 0{,}54 \; ou \; x > 0{,}66\}$.

11. Uma empresa não pode produzir mais que 5% de unidades defeituosas de um artigo num mesmo lote. Seja p a proporção de unidades defeituosas em um certo lote e suponha que, nesse lote, 100 artigos são sorteados para serem inspecionados. Responda as seguintes questões:

a. Qual o parâmetro que se deseja testar?

b. Qual é o estimador a ser utilizado e sua distribuição?

c. Indique as hipóteses a serem testadas e interprete-as.

d. Determine o critério de decisão com nível de significância de 5%.

e. Com o critério obtido, calcule a probabilidade de aceitar um lote com 7% de defeituosos.

f. Se forem observadas 10 unidades defeituosas, qual é o nível descritivo?

12. Uma urna contém bolas vermelhas e azuis. Para verificar a hipótese de iguais proporções dessas cores, extraem-se com reposição, 64 dessas bolas e decide-se aceitar a hipótese acima, se o número de bolas vermelhas retiradas estiver entre 28 e 36.

a. Determine a probabilidade de rejeitar a hipótese, quando ela é realmente correta.

b. Qual é a probabilidade do erro tipo II, se a verdadeira proporção de bolas vermelhas é 0,6?

c. Quanto vale a função poder, se a proporção de bolas vermelhas é 0,4?

13. A experiência mostra que a taxa de complicações, associada a um determinado procedimento cirúrgico, é de 0,20. Com o objetivo de reduzir essa taxa, um pesquisador desenvolveu um novo procedimento e o aplicou a uma amostra de pacientes.

a. Se ele usar a nova técnica em 100 pacientes, qual deveria ser a taxa limite para que conclua que a nova técnica é melhor que a anterior? Fixe o nível de significância em 0,05.

b. Se a verdadeira taxa de complicações associada à nova técnica for 0,08; qual é a probabilidade de que, em uma amostra de tamanho 100, ele não consiga rejeitar a hipótese nula?

c. Suponha que o pesquisador mantenha $\alpha = 0{,}05$ e deseje $\beta_{0,1} = 0{,}05$. Qual deve ser o tamanho da amostra para que isso aconteça?

14. Entre milhares de casos de pneumonia não tratados com sulfa, a porcentagem que desenvolveu complicações foi de 10%. Com o intuito de saber se o emprego das sulfas diminuiria essa porcentagem, 120 casos de pneumonia

foram tratados com sulfapiridina e destes, 6 apresentaram complicações. Admitindo que os pacientes são comparáveis em tudo, exceto quanto ao tratamento, teste a hipótese de que a proporção de casos com complicações entre os pacientes tratados com sulfa é significativamente menor do que os não tratados. Calcule o nível descritivo e tome a decisão considerando $\alpha = 0{,}05$.

15. Uma caixa tem bolas pretas e brancas. Existem 3 de uma cor e 2 de outra, mas não se sabe ao certo qual a cor predominante. Retiramos, ao acaso e com reposição, 50 bolas da urna e observamos 28 bolas brancas. Seja p a proporção de bolas brancas na caixa, pergunta-se:

a. Qual seria o teste adequado para decidir sobre a composição da caixa?

b. Qual conclusão, ao nível $\alpha = 10\%$?

c. Determine a probabilidade do erro tipo II.

16. Um milionário dá uma grande festa e resolve fazer uma brincadeira com seus convidados (que a essa altura já estavam bem animados...). Dentre os presentes, tidos como bons degustadores de vinho, sorteia 30 pessoas e oferece a cada uma dois copos de champanhe, numerados 1 e 2, e solicita que indiquem qual deles tem champanhe importada (é obrigatório escolher um único copo). Na verdade, os dois copos contém a mesma champanhe nacional! Deseja-se decidir se o "grupo" ainda é um bom provador de champanhe.

a. Indique como formular as hipóteses nula e alternativa para esse problema.

b. Que decisão você tomaria, ao nível de significância $\alpha = 0{,}08$ se 23 pessoas escolhessem o copo número 1?

c. E se 24 pessoas escolhessem o copo 2?

17. Um comerciante compra frutas para revenda e seu preço prevê no máximo 5% de frutas estragadas para que ele consiga algum lucro. Como não tem recursos para contratar um estatístico, ele adota a seguinte regra prática: toma, de cada lote, 30 frutas ao acaso. Se 3 ou mais estiverem estragadas, o lote é devolvido; caso contrário é aceito.

a. Qual a probabilidade do erro tipo I, no teste adotado pelo comerciante?

b. Esboce a função de poder do teste. Comente o teste do comerciante.

18. Um dado é lançado 216 vezes e o número de vezes que ocorreu a face 6 é contado. Decide-se aceitar a hipótese de que o dado é honesto, se o número de ocorrências estiver entre 31 e 41.

a. Formule as hipóteses nula e alternativa e indique a forma da região crítica.

b. Qual é a probabilidade do erro tipo I?

c. Qual seria a região crítica do teste ao nível 2% de significância?

19. Suponha que se deseje estimar a proporção p de indivíduos com certa moléstia em uma dada região. Selecionou-se uma amostra aleatória de 100 pessoas e constatou-se que 25 eram portadoras da moléstia.

a. Calcule a estimativa pontual da proporção p .

b. Construa um intervalo de confiança para p com coeficiente de confiança $\gamma = 0{,}95$. Qual o comprimento do intervalo?

c. Um pesquisador acredita que a proporção de doentes é superior a 20%. Teste essa hipótese ao nível $\alpha = 0{,}05$. Formule as hipóteses nula e alternativa.

20. Testes exaustivos realizados pela indústria *Cookbem* indicam que seu forno de microondas tem probabilidade 0,1 de apresentar a 1a. falha antes de 900 horas de uso. Um novo método de produção está sendo implantado e os engenheiros garantem que a probabilidade acima indicada deve diminuir. Com vistas a verificar essa afirmação, escolheu-se aleatoriamente 100 aparelhos para realizar testes acelerados e os resultados indicaram que 8 deles tiveram sua 1a. falha antes de 900 horas.

a. Formule as hipóteses adequadas.

b. Determine o nível descritivo.

c. Verifique se os engenheiros têm razão, considerando um nível de significância $\alpha = 6\%$.

21. Uma amostra de 10 adultos, na faixa de idade de 19 a 25 anos, apresentou uma freqüência cardíaca média de 68,7 batidas/min, com desvio-padrão de 8,67 batidas/min. Um manual de procedimento clínico indica que a pulsação média para indivíduos nessa faixa etária deve ser igual a 72 batidas/min. Admitindo que a variável medida se comporte de acordo com um modelo Normal e usando um nível de significância igual a $\alpha = 4\%$, você diria que os dados fornecidos são compatíveis com a informação do manual? Qual é o nível descritivo correspondente aos resultados fornecidos pela amostra?

22. A resistência à ruptura em cabos de aço é considerada uma variável Normal com média e variância dependendo de outros fatores. Uma amostra de 12 cabos produzidos por uma empresa são levados a teste para indicar se eles podem ser usados na construção de uma ponte. Cada cabo para ser usado precisa ter carga média de ruptura de no mínimo 2500 kg. Indique a conclusão que se pode tirar, baseado no nível descritivo, se os seguintes resultados forem observados na amostra: 2518, 2492, 2450, 2535, 2547, 2486, 2455, 2499, 2522, 2505, 2469 e 2440.

23. O crescimento de bebês, durante o primeiro mês de vida, pode ser modelado pela distribuição Normal. Admita que, em média, um crescimento de 5 centímetros ou mais seja considerado satisfatório. Deseja-se verificar se o crescimento de bebês de famílias em um bairro da periferia de São Paulo acompanha o padrão esperado. Para tanto, 10 recém-nascidos na região foram sorteados e sua altura acompanhada, fornecendo as seguintes medidas de crescimento em centímetros: 5,03; 5,02; 4,95; 4,96; 5,01; 4,97; 4,90; 4,91; 4,90 e 4,93.

a. Que hipóteses estão sendo testadas?

b. Qual é o estimador a ser utilizado para testar as hipóteses em (a) e qual é a sua distribuição?

c. Se a região crítica construída é $\{t \in \mathbb{R} : t \leq -3{,}87\}$, encontre o valor de α. Qual a conclusão?

d. Qual seria a região crítica e a conclusão se $\alpha = 5\%$?

24. A porcentagem anual média da receita municipal empregada em saneamento básico em pequenos municípios de um estado tem sido 8% (admita que esse índice se comporte segundo um modelo Normal). O governo pretende melhorar esse índice e, para isso, ofereceu alguns incentivos. Para verificar a eficácia dessa atitude, sorteou 10 cidades e observou as porcentagens investidas no último ano. Os resultados foram (em porcentagem) 8, 10, 9, 11, 8, 12, 16, 9, 11 e 12. Os dados trazem evidência de melhoria, ao nível de 2%? Caso altere a média, dê um intervalo de confiança para a nova média.

25. Alguns cientistas acreditam que em média 50% dos materiais expelidos por erupções vulcânicas são constituídos de enxofre. Seja X a massa de enxofre contida a cada 2 quilos de material vulcânico. Acredita-se que essa variável não tem distribuição Normal. Uma amostra de 100 caixas de 2 quilos desse material forneceu $\sum_{i=1}^{100} x_i = 98$ (em kg) e $\sum_{i=1}^{100} x_i^2 = 100$ (em kg^2).

a. Qual a distribuição de $\overline{X}$? Indique as suposições feitas.

b. Formule as hipóteses e obtenha a região crítica para $\alpha = 5\%$.

c. Qual a conclusão do teste?

d. Qual é a probabilidade do erro tipo II, se os vulcões expelem 52% de enxofre?

26. Deseja-se verificar se o modelo Uniforme Discreto com valores de 0 a 5 pode ser usado para modelar o número de reclamações que chegam por hora a uma Central de Atendimento ao Consumidor. O sorteio de 120 períodos de uma hora forneceu os seguintes dados:

Reclamações	0	1	2	3	4	5
Freqüência	8	32	28	24	16	12

Formule as hipóteses testadas e dê a conclusão ao nível de 5% de significância.

27. Uma indústria registra, em cada semana, o número de dias em que ocorrem acidentes de trabalho. Para uma amostra de 200 semanas, verifique se os dados apresentados a seguir, aderem ao modelo Binomial com parâmetros $n = 5$ e $p = 0{,}2$ (use nível de significância de 10%).

No. de dias com acidentes	0	1	2	3	4	5
Freqüência	64	56	40	24	8	8

28. O número de chegadas de clientes a um banco foi anotado minuto a minuto para uma amostra de 70 períodos (de um minuto). Os dados foram os seguintes:

No. Chegadas	0	1	2	3	4	5	6	mais de 6
Freqüência	9	15	17	11	7	5	4	2

O modelo de Poisson foi proposto para modelar essas chegadas, qual a sua opinião?

29. O tempo residual do efeito de um agrotóxico está sendo analisado. Estudos anteriores, com produtos similares, indicam que o modelo Exponencial com média de 3 dias poderia ser adequado. Qual a conclusão, ao nível 5%, se a análise em laboratório de uma amostra de 300 aplicações do agrotóxico forneceu os seguintes tempos residuais:

Faixas de Tempo	$[0,1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,\infty)$
Freqüência	89	60	43	40	25	22	21

30. O preço unitário de mudas de laranjeira (em reais), em atacadistas especializados, é uma variável aleatória que se pretende modelar pela Normal. Com base nos dados apresentados na tabela a seguir, teste a hipótese de que o

modelo Normal é adequado a esse caso, considerando um nível de significância de 5%.

Faixas de Preço	Freqüência
[0,50; 0,60)	23
[0,60; 0,65)	36
[0,65; 0,70)	64
[0,70; 0,75)	95
[0,75; 0,80)	102
[0,80; 0,85)	71
[0,85; 0,90)	45
[0,90; 1,00)	14

31. Usando os dados abaixo, verifique descritivamente se existe dependência entre o número de filhos e o rendimento familiar (em reais). Em seguida, teste tal fato considerando um nível de significância de 1%. (Não esqueça de formular as hipóteses).

Renda \ Filhos	0	1	2	> 2
menos de 2000	15	27	50	43
2000 a 5000	25	37	12	8
5000 ou mais	8	13	9	10

32. Em uma escola de ensino médio, o desempenho dos alunos em matemática e física foi observado (ver tabela a seguir) para testar se existe dependência entre as duas disciplinas.

Física \ Matemática	Notas Altas	Notas Regulares	Notas Baixas
Notas Altas	46	71	22
Notas Regulares	47	143	58
Notas Baixas	29	72	40

Calcule o nível descritivo. Qual a decisão, ao nível de significância 2% ?

33. Acredita-se que o empenho de estudantes universitários muda no decorrer do curso. Para investigar essa afirmação, decidiu-se estudar a relação entre ano de curso e aprovação em disciplinas. Os pesquisadores obtiveram os registros de 186 estudantes universitários, selecionados aleatoriamente, dentre a totalidade de alunos de uma certa instituição de ensino superior. Foram consideradas 3

disciplinas básicas em cada ano. Os resultados obtidos foram resumidos e são apresentados na tabela a seguir.

Aprovação\ Ano	1	2	3	4
Todas	6	5	7	10
Duas	10	16	19	18
Uma	23	20	15	7
Nenhuma	15	7	6	2

a. Quantifique o grau de associação entre aprovação e ano cursado.
b. Teste a hipótese de que as duas variáveis são independentes, ao nível de significância de 5%.
c. Obtenha o nível descritivo.

34. Quatro grupos de pacientes com úlcera duodenal foram submetidos a diferentes cirurgias caracterizadas pela porcentagem de tecido gástrico removido. A tabela apresentada a seguir contém dados referentes à classificação dos pacientes quanto à severidade de uma seqüela indesejável da cirurgia.

Cirurgia\ Seqüela	Nenhuma	Pouca	Moderada	Total
V+D (0%)	61	28	7	96
V+A (25%)	68	23	13	104
V+H (50%)	58	40	12	110
G+R (75%)	53	38	6	97

Verificar se existe associação entre a porcentagem de tecido gástrico removido e a severidade da seqüela. Utilize o nível descritivo.

35. Investiga-se, para um certo produto, a fidelidade (alta, média e baixa) de seus consumidores. Em uma amostra de 200 homens e 200 mulheres, foram classificados como tendo alto grau de fidelidade 120 homens e 80 mulheres, enquanto com grau médio, 50 mulheres e 50 homens. Os dados fornecem evidências (use α= 2%) de possíveis diferenças de grau de fidelidade entre os sexos? Indique o teste realizado.

36. Um levantamento inicial sugere que o número de filhos depende da renda familiar dos pais. Para confirmar essa suspeita, amostras de famílias foram coletadas, em cada classe social, e o número de filhos em cada família foi contado. Verifique através de um teste de hipóteses se a variável tem

comportamento diferente em cada uma das subpopulações estudadas (use nível de significância de 1%).

Classe\ Filhos	0	1	2	3	> 3
Baixa	15	27	40	64	54
Média	25	27	28	12	8
Alta	10	25	15	8	2

37. Deseja-se comparar o grau de instrução dos habitantes de quatro cidades brasileiras que têm aproximadamente o mesmo tamanho populacional. Uma amostra de 100 habitantes foi sorteada em cada cidade e o nível educacional das pessoas foi observado. Dos resultados apresentados na tabela abaixo, que conclusão podemos tirar, usando o nível descritivo?

Cidade\ Instrução	Fundamental	Médio	Superior	Pós Graduação
A	26	65	8	1
B	10	46	30	14
C	73	22	5	0
D	55	40	2	3

38. A reação ao tratamento por quimioterapia foi estudada em quatro grupos de pacientes com câncer. Retirou-se uma amostra de pacientes de cada grupo e classificou-se a reação em três categorias: pouca, média e alta. Teste, ao nível de 2%, se todos os tipos de câncer reagem da mesma maneira.

Câncer\ Reação	Pouca	Média	Alta	Total
Tipo I	51	33	16	100
Tipo II	58	29	13	100
Tipo III	48	42	30	120
Tipo IV	26	38	16	80

39. Um índice sobre qualidade de vida foi observado em uma amostra de 400 idosos. Os dados são apresentados a seguir.

Faixas	$[0,10)$	$[10,20)$	$[20,30)$	$[30,40)$	$[40,45)$	$[45,50)$
Freq.	7	15	32	55	48	60
Faixas	$[50,55)$	$[55,65)$	$[65,70)$	$[70,75)$	$[75,85)$	$[85,100]$
Freq.	55	56	28	20	18	6

a. Teste se a média do índice é ou não igual a 50.

b. Com base no item (a), verifique se o modelo Normal é adequado para este índice. Como ficaria sua resposta sem utilizar a informação do item (a)?

40. As tabelas a seguir contêm o número de pessoas segundo origem e opinião a respeito do aborto.

Masculino

Origem\ Opinião	A favor	Contra
Capital	10	45
Interior	18	90

Feminino

Origem\ Opinião	A favor	Contra
Capital	55	40
Interior	22	20

a. Para cada sexo, verifique se origem e opinião são independentes.

b. Combine as informações em uma única tabela desconsiderando sexo e teste novamente a independência das variáveis.

c. Discuta os resultados obtidos em (a) e (b).

41. (Use o computador) Considerando os dados do arquivo *cancer.txt* descrito no Exercício 24 do Capítulo 1, defina dois grupos: um de pacientes jovens, com idades inferiores a 54 anos, e um de pacientes idosos, com idades superiores a 54 anos. Os grupos deverão conter 191 e 171 pacientes. Considere a variável nitrogênio na uréia (N).

a. Construa um *box-plot* para a variável N, para cada um dos grupos etários e compare-os descritivamente. Com base nos gráficos, existem indicações de que a idade está influenciando a concentração de nitrogênio na uréia?

b. É de interesse verificar se a média populacional da variável N para os pacientes idosos é superior a 15. Supondo que o modelo Normal com desvio padrão $\sigma = 7$ é adequado, qual a conclusão que pode ser tirada, para um nível de significância $\alpha = 0{,}001$?

c. Considerando agora o grupo de pacientes mais jovens, verifique se a média populacional para N é menor que 15. Suponha que o desvio padrão populacional é conhecido igual a 5 e que o modelo Normal é adequado. Obtenha o nível descritivo.

d. Com base nos resultados dos itens (b) e (c), discuta o comportamento das médias da variável N para os dois grupos de pacientes.

42. (Use o computador) Suponha que os dados do arquivo *areas.txt* (ver descrição no Exercício 25, Capítulo 1) corresponde a uma amostra de vários empreendimentos de uma mesma empreiteira. Segundo o memorial descritivo do empreendimento, as unidades devem ter área total igual a 50 m^2, independentemente do bloco. Entretanto, suspeita-se que as unidades do bloco B não satisfazem a essa especificação.

a. Para cada bloco, construa um histograma para as áreas de cada um dos cômodos das unidades. Compare os gráficos. Para cada cômodo, discuta se existe diferença entre os tamanhos para unidades do bloco A quando comparadas com unidades do bloco B?

b. Teste a hipótese de que os apartamentos dos edifícios construídos satisfazem, em média, ao memorial descritivo que especifica que a área da sala deve ter no mínimo 28 m^2. Use $\alpha = 0{,}05$.

c. Construa uma nova variável com a informação da área total (isto é, obtenha uma variável que é a soma das áreas de cada cômodo) e faça histogramas considerando cada bloco. O modelo Normal lhe parece adequado para descrever o comportamento dessa nova variável?

d. Verifique se, em média, a área total para as unidades do bloco A são inferiores ao valor estipulado. Use $\alpha = 0{,}03$ e assuma que o modelo Normal é adequado.

e. Repita o item (d) para as unidades do bloco B.

f. Baseado nos resultados obtidos nos itens anteriores, você diria que a empreiteira está, de fato, desrespeitando a metragem estipulada?

43. (Use o computador) Considerando os dados do Exercício 40, suponha que alguns proprietários de unidades, vendidas pela empreiteira, estão movendo uma ação judicial. Eles alegam que a proporção de apartamentos apresentando rachaduras está exageradamente alta, superando 30% das unidades.

a. Utilizando a variável Rachadura, discuta quais modelos probabilísticos poderiam ser utilizados para descrevê-la.

b. Verifique se a proporção de apartamentos apresentando rachaduras está, de fato, muito alta. Utilize um nível de significância de 2%.

44. (Use o computador) Considere os dados do arquivo *aeusp.txt* (Exercício 26, Capítulo 1) que corresponde a uma amostra da população de baixa renda da região do Butantã-SP.

a. Teste se o número médio de residentes em casas da população acima mencionada é inferior a 4. Indique as suposições adicionais necessárias.

b. Verifique estatisticamente se a proporção de trabalhadores com carteira assinada, nessa população, é inferior a 40%. Use $\alpha = 5\%$.

45. (Use o computador) Continuando com os dados do arquivo *aeusp.txt* (ver descrição no Exercício 26, Capítulo 1), teste a independência entre as variáveis Comun e Renda. Tome sua decisão calculando o nível descritivo.

Capítulo 9

Tópicos Especiais

9.1 Introdução

Neste capítulo, desenvolveremos alguns tópicos complementares que têm grande utilidade no trabalho cotidiano de profissionais e pesquisadores que fazem uso de métodos quantitativos. Boa parte da teoria necessária ao desenvolvimento das técnicas a serem apresentadas já foi discutida nos capítulos precedentes. Alguns dos tópicos podem ser elaborados em maiores detalhes e generalidade, entretanto, esta abordagem está além dos objetivos deste texto e demandaria capítulos específicos. Dessa forma, para esses casos apresentaremos uma versão mais simples a título de introdução ao tema. Para um aprofundamento maior, recomendamos consultar a bibliografia apresentada ao final do livro.

Num primeiro momento, consideraremos a comparação de duas médias populacionais. Em seguida, apresentaremos alguns procedimentos destinados ao estudo da variância populacional, incluindo aí a comparação de variâncias de duas populações. Uma generalização do método de comparação de médias, envolvendo mais que duas populações, é considerada no tópico seguinte. O desenvolvimento é baseado no conceito de *modelo estatístico*, que consiste em decompor o valor de cada observação em uma parte com estrutura conhecida e uma parte residual, com informações que não são explicadas pela mencionada estrutura. Quando o modelo estatístico é expresso em termos da equação de uma reta, podemos definir um modelo de regressão linear simples, que será discutido na última seção.

9.2 Comparação de Duas Médias

Considere que estamos interessados em comparar duas populações com relação às suas médias. Uma das principais suposições feita no desenvolvimento dos testes de hipóteses, apresentados anteriormente, foi a de independência entre os componentes da amostra. Ao tratarmos da comparação de parâmetros de duas populações, precisamos verificar se estas estão ou não relacionadas. Se estiverem

relacionadas, pode ocorrer dependência entre elementos de amostras diferentes, ainda que cada amostra, internamente, seja composta por elementos independentes. Por exemplo, é muito comum a situação em que observações são tomadas em uma mesma unidade amostral, antes e depois de alguma intervenção. Neste caso é razoável considerarmos que a repetição de mensurações em um mesmo elemento cause dependência entre as duas amostras. Veremos, nos próximos exemplos, alguns casos com e sem a presença de independência.

Exemplo 9.1: Uma distribuidora de combustíveis deseja verificar se um novo tipo de gasolina é eficaz na revitalização de motores velhos. Com esse objetivo, seleciona 12 automóveis de um mesmo modelo com mais de 8 anos de uso e, após regulagem de seus motores, verifica o consumo de combustível. Em seguida, o carro é abastecido com o novo tipo de combustível durante 15 semanas, e uma nova aferição do consumo é feita. Defina as variáveis aleatórias X_i e Y_i como o rendimento do automóvel i respectivamente *antes* e *após* as 15 semanas. Vemos que X_i e Y_i foram medidas em uma mesma unidade amostral e, assim, é razoável assumir que exista alguma dependência entre elas. Ressaltamos que, para $i \neq j$, devemos ter X_i e X_j independentes. O mesmo deve ocorrer para Y_i e Y_j. Ao medir a característica de interesse em duas ocasiões, para cada uma das unidades amostrais, pretende-se diminuir a influência de outros fatores (muitas vezes impossíveis de serem controlados) e ressaltar um possível efeito do tipo de gasolina no desempenho do veículo. □

Exemplo 9.2: Um estudo envolve a avaliação de um novo sistema operacional de computador, desenvolvido para crianças com idades entre 8 e 12 anos. Afirma-se que o novo sistema é mais rápido do que o atual, líder de mercado. Para testar esta afirmação, foram selecionados em uma mesma escola dois grupos com 15 crianças cada. As crianças, sem conhecimento prévio relacionado ao uso de computadores, utilizaram máquinas de mesma configuração para realizar uma certa tarefa, que teve seu tempo anotado. O primeiro grupo, denominado Grupo A, trabalhou com o sistema operacional convencional ao passo que o segundo grupo, Grupo B, desenvolveu atividades no novo sistema. Ao final do experimento todas as 30 crianças haviam realizado a tarefa. Nesse exemplo, os dois grupos selecionados consistem de 15 crianças diferentes e, portanto, pode-se assumir que os dois grupos constituem duas amostras independentes. □

A independência ou não das observações é um fator importante a ser considerado mas, como visto em capítulos anteriores, também é importante levarmos em consideração a variabilidade associada aos valores populacionais e amostrais. Note que, nos testes de média, utilizamos o valor da variância

populacional ou um estimador apropriado. Dessa forma, para procedermos à comparação das médias, precisamos analisar o que ocorre com as variâncias nas duas populações. Em algumas situações elas são conhecidas por estudos anteriores, censos ou ainda suposições. Se as variâncias populacionais são desconhecidas, existe ainda a questão delas serem iguais ou diferentes. Alguns autores argumentam que, no caso de variabilidades populacionais desiguais, o teste de médias não deveria ser realizado pois as populações já são diferentes. Apesar dessas opiniões, iremos apresentar neste capítulo um procedimento para esse caso. Resumimos, a seguir, as possíveis situações na comparação de duas populações.

Figura 9.1: Casos na comparação de duas amostras.

Discutiremos apenas os testes conhecidos como paramétricos, para cada um dos casos considerados na Figura 9.1. Os testes paramétricos assumem que as variáveis se comportam segundo um modelo Normal, ou que as amostras são suficientemente grandes, de modo que uma boa aproximação pode ser conseguida utilizando o modelo Normal.

Caso 1: Amostras dependentes (teste t - pareado)

No caso de amostras dependentes, desejamos comparar duas médias populacionais sendo que, para cada unidade amostral, realizamos duas medições da característica de interesse. De modo geral, essas observações correspondem a

medidas tomadas antes e após uma dada intervenção. Para ilustrar esta situação, considere o Exemplo 9.1 em que medimos o rendimento com a gasolina tradicional e depois com o novo tipo de combustível. Essa é uma típica situação em que o teste t - pareado deve ser utilizado. Neste caso, é de se esperar que exista alguma correlação entre as observações tomadas em uma mesma unidade experimental.

As medidas tomadas antes e após a intervenção realizada serão representadas pelas variáveis aleatórias X_i e Y_i, respectivamente. Desta forma, o efeito produzido pode ser representado, para o i-ésimo indivíduo, pela variável $D_i = Y_i - X_i$. Supondo, para $i = 1, \ldots, n$,

$$D_i \sim N(\mu_D, \sigma_D^2),$$

queremos testar as hipóteses:

$$H_o : \mu_D = 0 \text{ (a intervenção não produz efeito)}$$
$$H_a : \mu_D \neq 0 \text{ (a intervenção produziu algum efeito)},$$

sendo que a hipótese alternativa pode também ser unilateral.

O parâmetro μ_D é estimado pela média amostral $\overline{D}$ e, como usualmente não temos informação sobre σ_D^2, estimamos seu valor por S_D^2, dado por

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2.$$

O teste de hipóteses é realizado utilizando-se a quantidade

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}},$$

que, sob H_o, segue uma distribuição t-Student com $n-1$ graus de liberdade. O teste segue os mesmos passos discutidos no capítulo anterior.

Exemplo 9.3: No Exemplo 9.1, o rendimento foi representado por X_i e Y_i para o automóvel i, respectivamente antes e após o novo combustível. Os valores observados, em km/l, junto com as diferenças $D_i = Y_i - X_i$, para os 12 automóveis são apresentados na tabela a seguir.

Autom.	1	2	3	4	5	6	7	8	9	10	11	12
Após (Y)	11,6	8,8	9,9	9,5	11,6	9,1	10,6	10,8	13,4	10,6	10,5	11,4
Antes (X)	8,1	7,9	6,8	7,8	7,6	7,9	5,7	8,4	8,0	9,5	8,0	6,8
$D = Y - X$	3,5	0,9	3,1	1,7	4,0	1,2	4,9	2,4	5,4	1,1	2,5	4,6

Os dados de consumo, antes a após o novo combustível, podem ser visualizados através de gráficos *box-plot*, apresentados a seguir, de onde podemos notar indicações de que o rendimento é, aparentemente, maior após o uso do novo combustível.

Para podermos verificar se o rendimento é de fato superior, precisamos proceder ao teste:

$$H_o : \mu_D = 0 \text{ (o novo combustível não aumenta o rendimento)};$$
$$H_a : \mu_D > 0 \text{ (o novo combustível aumenta o rendimento)},$$

com μ_D representando o valor esperado da diferença de rendimento, isto é, $\mu_D = E(Y - X)$. Estaremos assumindo que a distribuição de $D_i = Y_i - X_i$, para $i = 1, \ldots, 12$, é Normal com média μ_D e variância σ_D^2.

Com os dados observados, obtemos $\overline{d}_{obs} = 2{,}9$ e estimamos σ_D^2 por $s_{D_{obs}}^2 = 2{,}4$. Logo, sob H_o,

$$t_{obs} = \frac{\overline{d}_{obs} - \mu_D}{s_{D_{obs}}/\sqrt{n}} = \frac{2{,}9 - 0}{1{,}55/\sqrt{12}} = 6{,}48.$$

Com $\alpha = 0{,}05$ e utilizando a tabela da distribuição t-Student com 11 graus de liberdade, obtemos t_c resolvendo a equação $P(T > t_c) = 0{,}05$. Obtemos $t_c = 1{,}796$ e como $t_{obs} > t_c$, concluímos que o novo combustível é eficaz na

melhora do rendimento, acarretando diminuição do consumo para o tipo de veículo considerado no experimento. □

Caso 2. Amostras independentes com variâncias conhecidas

Consideramos agora o teste relacionado com a situação em que queremos comparar médias de duas populações independentes, quando as correspondentes variâncias são conhecidas. A obtenção de informação a respeito do valor da variância populacional pode ser obtido de estudos anteriores ou experimentos similares.

Exemplo 9.4: Vimos no Exemplo 9.2 que, para comparar dois sistemas operacionais, dois grupos independentes de estudantes foram selecionados e o tempo necessário para realizar a tarefa foi anotado.

Os dados obtidos foram os seguintes (em minutos):

Grupo	Tempo														
A	182	185	193	175	184	192	175	173	186	178	162	179	164	182	186
B	92	76	76	90	97	90	86	93	100	115	85	80	90	86	94

A inspeção visual dos dados sugere que o Grupo B tende a realizar a tarefa num tempo inferior àquele observado para o Grupo A. Para auxiliar esta análise inicial, podemos construir gráficos *box-plot* para os grupos e colocá-los lado a lado conforme a figura a seguir.

Podemos observar que, para os alunos considerados, o novo sistema operacional oferece maior facilidade de aprendizado, caracterizado aqui pelo tempo de execução de certa tarefa, uma vez que o *box-plot* para o Grupo B está sensivelmente mais baixo.

Note que o valor da mediana do Grupo B é inferior ao do Grupo A, mas o intervalo entre o primeiro e o terceiro quartil é próximo para os dois grupos, dando a idéia de que a variabilidade do tempo de aprendizado é semelhante para ambos os sistemas operacionais.

É importante ressaltar que, para podermos concluir que o novo sistema é de fato eficaz, precisamos extrapolar as conclusões anteriores para toda a população de crianças com idade entre 8 e 12 anos. Isto pode ser feito, realizando o teste de hipóteses que será descrito em seguida a este exemplo.

Outras medidas descritivas podem ser calculadas para auxiliar na comparação. □

Utilizando a motivação fornecida pelo exemplo anterior, podemos considerar a seguinte situação geral. Suponha que desejamos comparar duas populações, cujas variâncias são iguais a um valor conhecido σ_o^2. Além disso, vamos admitir que estas duas populações se comportam conforme um modelo Normal, com médias μ_1 e μ_2. Sejam X e Y as variáveis aleatórias representando a característica de interesse em cada uma das populações. Segue, portanto, que os dados poderão ser representados por dois conjuntos de variáveis aleatórias independentes $(X_1, \ldots, X_{n_1})$ e $(Y_1, \ldots, Y_{n_2})$, representando amostras aleatórias, obtidas de cada uma das populações. Deve ser notado que os tamanhos de amostra n_1 e n_2 podem, eventualmente, ser iguais. Queremos testar

H_o : As médias populacionais são iguais;
H_a : As médias populacionais não são iguais.

Estas hipóteses podem ser traduzidas em termos de μ_1 e μ_2:

$$H_o : \mu_1 = \mu_2;$$
$$H_a : \mu_1 \neq \mu_2.$$

Se a suspeita sobre a diferença entre as médias é de que a média de uma população é maior (ou menor) do que a média da outra, podemos reescrever H_a como $\mu_1 > \mu_2$ (ou $\mu_1 < \mu_2$) e proceder ao teste unilateral.

Como estamos interessados em determinar se a diferença é estatisticamente significante, podemos ainda reescrever as hipóteses em termos de $\mu_D = \mu_1 - \mu_2$, isto é,

$$H_o : \mu_D = 0;$$
$$H_a : \mu_D \neq 0,$$

o que sugere trabalharmos com o *estimador* de μ_D:

$$\overline{D} = \overline{X} - \overline{Y}.$$

Com as suposições feitas, temos

$$X_i \sim N(\mu_1, \sigma_o^2),\ i = 1, 2, ..., n_1;$$
$$Y_i \sim N(\mu_2, \sigma_o^2),\ i = 1, 2, ..., n_2\,.$$

Pela independência dessas variáveis, $\overline{D}$ terá distribuição Normal com média $E(\overline{D}) = \mu_D$ e quanto à variância, temos:

$$\begin{aligned} Var(\overline{D}) &= Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + Var(\overline{Y}) \\ &= \frac{\sigma_o^2}{n_1} + \frac{\sigma_o^2}{n_2} = \sigma_o^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \end{aligned}$$

Note que a independência entre as amostras foi necessária para obter essa variância, uma vez que a covariância entre as médias amostrais é zero.

Com estas informações, procedemos ao teste de hipóteses do modo usual. Caso não saibamos qual é a distribuição da característica na população podemos, para amostras de tamanho grande, lançar mão do Teorema Central do Limite e trabalhar, de modo aproximado, com a distribuição Normal.

Exemplo 9.5: Continuando o Exemplo 9.4, sejam T_1 e T_2 variáveis aleatórias representando os tempos de aprendizado para os grupos A e B, respectivamente.

Tendo em vista que $n_1 = n_2 = 15$, as amostras das respectivas populações são os conjuntos de variáveis aleatórias independentes $(T_{1,1}, \ldots, T_{1,15})$ e $(T_{2,1}, \ldots, T_{2,15})$. Além disso, assuma que informações adicionais fornecidas pelas empresas indicam que a variabilidade dos tempos de aprendizado é a mesma para ambos os sistemas operacionais e igual a $\sigma_T = 10$ min. Logo, para $i = 1, 2, \ldots, 15$,

$$T_{1i} \sim N(\mu_1, 100);$$
$$T_{2i} \sim N(\mu_2, 100).$$

Queremos testar

H_o : Tempo médio é igual para ambos os sistemas ;
H_a : Aprendizado do novo sistema é, em média, mais rápido.

As hipóteses podem ser formuladas como

$$H_o : \mu_1 = \mu_2;$$
$$H_a : \mu_1 > \mu_2,$$

ou, equivalentemente,

$$H_o : \mu_D = \mu_1 - \mu_2 = 0;$$
$$H_a : \mu_D = \mu_1 - \mu_2 > 0.$$

A região crítica será dada por $RC = \{d \in \mathbb{R} : d > d_c\}$ e o estimador de μ_D será dado por $\overline{D} = \overline{T}_1 - \overline{T}_2$, com

$$\overline{T}_1 = \frac{\sum_{i=1}^{15} T_{1,i}}{15} \quad \text{e} \quad \overline{T}_2 = \frac{\sum_{i=1}^{15} T_{2,i}}{15}.$$

Pela suposição de que os tempos seguem o modelo Normal e, lembrando que as amostras são independentes, segue que a distribuição de $\overline{D}$ é Normal com média μ_D e variância

$$Var(\overline{D}) = Var(\overline{T}_1) + Var(\overline{T}_2) = \frac{100}{15} + \frac{100}{15} = \frac{200}{15} = 13{,}33.$$

Utilizamos agora o procedimento usual para testes de hipóteses, fixando $\alpha = 0{,}05$ e encontrando um valor crítico d_c tal que

$$\begin{aligned} &P(\text{rejeitar } H_o \mid H_o \text{ verdadeira}) = P(\overline{D} \in RC \mid \mu_D = 0) \\ &= P\left(\frac{\overline{D} - \mu_D}{\sqrt{13{,}33}} > \frac{d_c - 0}{\sqrt{13{,}33}}\right) = P(Z > \frac{d_c}{3{,}65}) = 0{,}05\,. \end{aligned}$$

Consultando a tabela da distribuição Normal padrão, obtemos $z_c = 1{,}64$. Logo,

$$d_c = 1{,}64 \times 3{,}65 = 5{,}99.$$

Então, $RC = \{d \in \mathbb{R} \,|\, d > 5{,}99\}$.

Considerando os valores amostrais observados, temos que a média para o grupo A é 179,73 min e, para o grupo B, é de 89,86 min. Assim,

$$\overline{d}_{obs} = 179{,}73 - 89{,}86 = 89{,}87.$$

Como $\overline{d}_{obs} \in RC$, rejeitamos a hipótese nula, isto é, a um nível de significância de 5% concluímos que, para alunos com idade entre 8 e 12 anos sem conhecimento computacional prévio, o tempo de aprendizado com o novo sistema operacional é menor. □

Consideramos, agora, a situação em que as populações apresentam médias desconhecidas e variâncias populacionais conhecidas, porém com valores diferentes. Nesse caso, já sabemos que as populações são diferentes, uma vez que as variabilidades da característica de interesse nas duas populações são diferentes. Ainda assim, podemos estar interessados em verificar se as médias também são diferentes e utilizar a teoria de teste de hipóteses, para embasar estatisticamente a decisão a ser tomada.

Com as suposições e a notação já apresentada anteriormente, temos agora que $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$, com $\sigma_X \neq \sigma_Y$. Então,

$$\overline{X} \sim N(\mu_X, \sigma_X^2/n_1) \quad \text{e} \quad \overline{Y} \sim N(\mu_Y, \sigma_Y^2/n_2)\,.$$

Para $\overline{D} = \overline{X} - \overline{Y}$ e utilizando a independência entre $\overline{X}$ e $\overline{Y}$, temos que

$$Var(\overline{D}) = Var(\overline{X}) + Var(\overline{Y}) = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2},$$

e, então, $\overline{D} \sim N(\mu_X - \mu_Y, \sigma_X^2/n_1 + \sigma_Y^2/n_2)$. A partir daqui, o teste prossegue na forma usual. No próximo exemplo, ilustramos o procedimento apresentado, de variâncias conhecidas porém diferentes.

Exemplo 9.6: Uma empresa avaliadora de imóveis está estudando as regiões central e oeste da cidade de São Paulo. O objetivo principal é verificar se o preço médio, praticado para imóveis comerciais de um dado tamanho, é o mesmo nas duas áreas. De levantamentos anteriores, a empresa sabe que a área oeste apresenta uma heterogeneidade de preços imobiliários (em UPC- unidade padrão de construção) maior do que a região central, sendo os desvios padrões iguais a 0,82 UPC para a região oeste e 0,71 UPC para a região central. Para verificar se os preços médios são iguais ou não, duas amostras, uma de tamanho 20 e outra de tamanho 18 foram retiradas aleatoriamente de cada região. Os dados são os seguintes:

Região Central

41,2	40,5	39,6	39,4	38,9	39,1	40,9	41,2	40,4	40,0
40,6	40,3	39,2	40,6	39,7	40,3	40,9	39,6	39,7	41,2

Região Oeste

37,2	34,9	38,1	35,4	35,7	37,7	36,4	36,6	36,1
37,4	36,1	35,9	36,9	37,4	37,5	38,0	36,8	36,4

Algumas medidas resumo são apresentadas na próxima tabela:

Medidas	Região		
Descritivas	Central	Oeste	Ambas
n	20	18	38
Média	40,2	36,7	38,5
Mediana	40,3	36,7	39,0
Desvio-Padrão	0,7	0,9	1,9
Mínimo	38,9	34,9	34,9
Máximo	41,2	38,0	41,2

O comportamento dos dados pode ser visualizado através de gráficos tipo *box-plot*, mostrados a seguir.

Note que o valor do desvio padrão amostral sugere, de fato, que as variâncias são diferentes nas duas regiões; mais ainda, a média de preço na região central parece ser superior à da região oeste. Para os dados observados, a região central tem, aparentemente, preços superiores à região oeste. Além disso, a variabilidade observada nos imóveis da região oeste é maior, o que, de certa forma confirma a informação fornecida pela empresa. Em resumo, para os dados apresentados nas duas amostras, temos um maior preço médio (amostral) para a região central. Essas conclusões são válidas apenas para os valores amostrais observados. Para podermos extrapolar esta conclusão para as regiões como um todo, precisaremos utilizar um procedimento estatístico que controle os erros, eventualmente, cometidos.

Representando a informação dos preços na região central pela variável aleatória X e, para a região oeste, pela variável aleatória Y, assumimos que os dados são obtidos de duas populações Normais de tal forma que

$$\overline{X} \sim N(\mu_X, \sigma_X^2/20) \quad \text{e} \quad \overline{Y} \sim N(\mu_Y, \sigma_Y^2/18).$$

Nosso principal interesse é testar as hipóteses

$$H_o : \mu_X = \mu_Y;$$
$$H_a : \mu_X \neq \mu_Y.$$

Definindo $\overline{D} = \overline{X} - \overline{Y}$ temos

$$Var(\overline{D}) = Var(\overline{X}) + Var(\overline{Y}) = \frac{0{,}71^2}{20} + \frac{0{,}82^2}{18} = 0{,}06.$$

Logo, para $\alpha = 0{,}05$ vem:

$$P(\text{rejeitar } H_o \mid H_o \text{ verdadeira}) = P(\overline{D} \in RC \mid \mu_X - \mu_Y = 0)$$
$$= P(Z < \frac{d_{c_1}}{\sqrt{0{,}06}} \; ou \; Z > \frac{d_{c_2}}{\sqrt{0{,}06}}) = 0{,}05.$$

Da tabela da distribuição Normal padrão obtemos os valores críticos:

$$\frac{d_{c_1}}{0{,}25} = -1{,}96 \Rightarrow d_{c_1} = -0{,}49;$$
$$\frac{d_{c_2}}{0{,}25} = 1{,}96 \Rightarrow d_{c_2} = 0{,}49.$$

Consequentemente,

$$RC = \{d \in \mathbb{R} : d < -0{,}49 \; ou \; d > 0{,}49\}$$

Como em nosso caso $\overline{d}_{obs} = 40{,}2 - 36{,}7 = 3{,}50$ pertence à região crítica, concluímos que os imóveis situados nas regiões central e oeste têm preços médios diferentes, ao nível de significância de 5%. □

Caso 3A: Amostras independentes com variâncias desconhecidas e iguais

No caso anterior vimos que informações adicionais podem fornecer subsídios para o conhecimento dos valores das variâncias populacionais. Em geral, contudo, não temos informações a respeito do valor das variâncias. Entretanto, os processos que geram os dados podem nos levar a crer que, apesar de desconhecidas, as variâncias são iguais para as duas populações.

Exemplo 9.7: Digitadores são treinados em uma empresa em duas turmas distintas. Na primeira, denominada Turma J, utiliza-se um método japonês de ensino, ao passo que na segunda turma, denominada Turma A, utiliza-se um método alemão. Deseja-se comparar os dois métodos e para tanto, 16 alunos de cada turma foram escolhidos aleatoriamente e uma mesma tarefa foi atribuída a cada um. Ao final do experimento, o tempo gasto na realização da tarefa, para cada aluno, foi anotado. No processo, dois computadores utilizados pelos alunos selecionados da turma J e três da turma A apresentaram problemas que impediram a realização da tarefa; o tamanho da amostra foi assim reduzido para 14 e 13, respectivamente, para as turmas J e A.

Os dados obtidos foram:

Turma	Tempos (min)													
J	10	13	9	10	14	13	10	15	12	10	9	10	13	14
A	15	12	18	16	15	17	17	15	16	17	11	17	14	

Apesar de não conhecidas, as variâncias populacionais para as duas turmas são consideradas iguais com base em estudos anteriores. □

Para formalizar a situação apresentada, supomos que os dados para o primeiro grupo são representados por variáveis aleatórias independentes $X_1, \ldots, X_{n_1}$ e, para o segundo, $Y_1, \ldots, Y_{n_2}$. Além disso, assumimos que

$$X_i \sim N(\mu_X, \sigma^2), \; i = 1, \ldots, n_1;$$
$$Y_j \sim N(\mu_Y, \sigma^2), \; j = 1, \ldots, n_2.$$

Para ambas as populações, temos a mesma variância σ^2 (desconhecida). Suponha que nosso interesse é testar

$$H_o : \mu_X = \mu_Y;$$
$$H_a : \mu_X \neq \mu_Y.$$

Novamente, consideramos o estimador $\overline{D}$ definido pela diferença $\overline{X} - \overline{Y}$. Dada a independência entre as amostras, segue imediatamente que

$$E(\overline{D}) = \mu_X - \mu_Y;$$
$$Var(\overline{D}) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Além disso, considerando também a normalidade dos dados, segue que

$$\overline{D} \sim N(\mu_X - \mu_Y, \sigma^2(1/n_1 + 1/n_2)).$$

e consequentemente,

$$\frac{\overline{D} - (\mu_X - \mu_Y)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1).$$

Como a variância populacional σ^2 é desconhecida, precisará ser estimada. Tendo em vista que S_X^2 e S_Y^2 são ambos estimadores não viciados dessa variância, usaremos como estimativa para σ^2 uma combinação deles, dada por:

$$S_c^2 = \frac{(n_1 - 1)\, S_X^2 + (n_2 - 1) S_Y^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{j=1}^{n_2}(Y_j - \overline{Y})^2}{n_1 + n_2 - 2}.$$

Note que S_c^2 é uma média ponderada entre S_X^2 e S_Y^2, com ponderação dada por $n_1 - 1$ e $n_2 - 1$. Dessa forma, estaremos utilizando para estimar σ^2, toda a informação disponível nas duas amostras. Além disso, pode-se mostrar que S_c^2 é não viciado para σ^2.

Da mesma forma que na Seção 8.3 do Capítulo 8, o uso do estimador S_c^2 nos leva a trabalhar com a distribuição t-Student, isto é,

$$T = \frac{\overline{D} - (\mu_X - \mu_Y)}{S_c\sqrt{1/n_1 + 1/n_2}}$$

tem, sob H_o, distribuição t-Student com $n_1 + n_2 - 2$ graus de liberdade.

Dada a hipótese alternativa apresentada, procedemos ao teste bilateral da forma usual, isto é, fixado α encontra-se o valor t_c tal que

$$\begin{aligned}\alpha &= P(\text{rejeitar } H_o \mid H_o \text{ verdadeira})\\ &= P(T < -t_c \ ou \ T > t_c \mid H_o).\end{aligned}$$

A quantidade t_c é então obtida da tabela da distribuição t-Student, com $n_1 + n_2 - 2$ graus de liberdade. A região crítica para o teste é dada por

$$RC = \{t \in \mathbb{R} : t < -t_c \ ou \ t > t_c\}.$$

Uma vez obtidas as amostras, substituindo as estimativas de $\overline{D}$ e S_c na expressão de T, obtemos o valor t_{obs}. Rejeitamos H_o se t_{obs} pertencer à região crítica.

Exemplo 9.8: Para o Exemplo 9.7, podemos escrever as hipóteses de interesse como

$$H_o : \mu_X = \mu_Y \text{ (os dois métodos são equivalentes)};$$
$$H_a : \mu_X \neq \mu_Y,$$

com μ_X e μ_Y representando, respectivamente, o tempo médio populacional para alunos da turma J e da turma A. As amostras forneceram os seguintes valores:

$$n_1 = 14, \overline{x}_{obs} = 11{,}57 \text{ e } s^2_{X_{obs}} = 4{,}1;$$
$$n_2 = 13, \overline{y}_{obs} = 15{,}38 \text{ e } s^2_{Y_{obs}} = 4{,}3.$$

Então,

$$\overline{d}_{obs} = \overline{x}_{obs} - \overline{y}_{obs} = 11{,}57 - 15{,}38 = -3{,}81;$$
$$s^2_{c_{obs}} = \frac{(n_1 - 1)\, s^2_{X_{obs}} + (n_2 - 1) s^2_{Y_{obs}}}{(n_1 - 1) + (n_2 - 1)} = \frac{13 \times 4{,}1 + 12 \times 4{,}3}{25} = 4{,}2.$$

Como a hipótese alternativa apresentada é bilateral, a região crítica tem a forma $RC = \{t \in \mathbb{R} : t < -t_c \ ou \ t > t_c\}$. Logo, para $\alpha = 0{,}01$ temos

$$\begin{aligned}0{,}01 &= P(\text{rejeitar } H_o \mid H_o \text{ verdadeira})\\ &= P(T < -t_c \ ou \ T > t_c \mid H_o).\end{aligned}$$

Da tabela da distribuição t-Student com 25 graus de liberdade, obtemos $t_c = 2{,}79$. Conseqüentemente,

$$RC = \{t \in \mathbb{R} : t < -2{,}79 \ \ ou \ \ t > 2{,}79\}.$$

Utilizando as estimativas calculadas temos, sob H_o,

$$t_{obs} = \frac{\overline{d}_{obs}}{\sqrt{s^2_{c_{obs}}(1/n_1 + 1/n_2)}} = \frac{-3{,}81}{\sqrt{4{,}2(1/14 + 1/13)}} = -4{,}83\,;$$

que pertence à região crítica e, assim, concluímos que os métodos de fato diferem, a um nível de significância de 1%. □

Caso 3B: Amostras independentes com variâncias desconhecidas e diferentes

O teste para o caso em que as variâncias são desconhecidas e desiguais é teoricamente mais envolvente. Assim, sem entrar em maiores detalhes, consideramos as mesmas hipóteses apresentadas no Caso 3A, só que, agora, a quantidade a ser usada para o teste será

$$T = \frac{\overline{D} - (\mu_X - \mu_Y)}{\sqrt{S^2_X/n_1 + S^2_Y/n_2}}\,.$$

A exemplo do caso anterior, t também tem distribuição t-Student, mas os graus de liberdade ν são corrigidos pela expressão

$$\nu = \frac{\left(S^2_X/n_1 + S^2_Y/n_2\right)^2}{\frac{\left(S^2_X/n_1\right)^2}{n_1-1} + \frac{\left(S^2_Y/n_2\right)^2}{n_2-1}}\,.$$

A seqüência do teste é similar àquela apresentada nos casos anteriores.

Na Tabela 9.1 mostramos um resumo dos testes considerados nesta seção.

Encerramos esta seção, considerando a situação em que a característica de interesse não se comporta segundo um modelo Normal. Novamente, a alternativa será coletar uma amostra de tamanho grande o suficiente, a fim de utilizar o Teorema Central do Limite e obter distribuições amostrais aproximadamente Normais. Como um exemplo desse procedimento, vamos desenvolver o teste para a igualdade de duas proporções.

Tabela 9.1: Comparação de médias para duas populações.

Situação	Estimadores
Amostras Pareadas (Caso 1)	$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$ $S^2_D = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$ $\frac{\overline{D} - \mu_D}{\sqrt{S^2_D/n}} \sim t_{(n-1)}$
Amostras Independentes (Caso 2) Variâncias conhecidas	$\overline{D} = \overline{X} - \overline{Y}$ $Var(\overline{D}) = \sigma^2_X/n_1 + \sigma^2_Y/n_2$ $\frac{\overline{D} - \mu_D}{\sqrt{\sigma^2_X/n_1 + \sigma^2_Y/n_2}} \sim N(0,1)$
Amostras Independentes (Caso 3A) Variâncias desconhecidas e iguais	$\overline{D} = \overline{X} - \overline{Y}$ $S^2_c = \frac{(n_1-1)\,S^2_X + (n_2-1)S^2_Y}{(n_1-1)+(n_2-1)}$ $\frac{\overline{D} - \mu_D}{\sqrt{S^2_c(1/n_1 + 1/n_2)}} \sim t_{(n_1+n_2-2)}$
Amostras Independentes (Caso 3B) Variâncias desconhecidas e diferentes (Para os graus de liberdade ν ver expressão no texto)	$\overline{D} = \overline{X} - \overline{Y}$ $\widehat{\sigma}^2_D = S^2_X/n_1 + S^2_Y/n_2$ $\frac{\overline{D} - \mu_D}{\sqrt{S^2_X/n_1 + S^2_Y/n_2}} \sim t_{(\nu)}$

Exemplo 9.9: Num estudo sobre doenças infantis, desejamos investigar se a incidência de casos de contaminação por vermes é afetada pela idade. Dois grupos de crianças, um com idades de 2 a 4 anos (Grupo I) e outro, com idades de 7 a 9 anos (Grupo II) foram escolhidos para serem examinados quanto à ocorrência de vermes. Os dados são apresentados a seguir:

Grupo	Amostra	Proporção com Verminose
I	120	0,085
II	260	0,103

Para saber se as duas faixas etárias acima têm o mesmo comportamento, quanto a incidência dessa doença, podemos realizar um teste de hipóteses envolvendo proporções. □

Considere que desejamos verificar o comportamento de uma certa característica em duas populações. Se a amostra for suficientemente grande sabemos, pelo Teorema Central do Limite, que a distribuição de probabilidade da proporção amostral tem um comportamento aproximadamente igual ao modelo Normal. Na comparação de proporções em duas populações, usaremos como estimador a diferença entre as respectivas proporções amostrais. Não é difícil verificar que ela será um estimador não viesado da respectiva diferença entre as proporções populacionais.

Supondo que duas amostras independentes foram retiradas, uma em cada população, teremos duas proporções amostrais independentes e a diferença entre elas também terá distribuição aproximadamente Normal. Assim, se o interesse é testar:

$$H_o : p_1 = p_2 \ \ versus \ \ H_a : p_1 \neq p_2 \,,$$

então o estimador a ser utilizado será $\widehat{p}_1 - \widehat{p}_2$, cuja distribuição será aproximada pela Normal cujos parâmetros são obtidos, considerando-se as relações:

$$E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2;$$

$$Var(\widehat{p}_1 - \widehat{p}_2) = Var(\widehat{p}_1) + Var(\widehat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\,.$$

Note que, para calcular a variância, a independência entre as amostras garantiu a independência entre $\widehat{p}_1$ e $\widehat{p}_2$ e, portanto, a covariância entre eles se anulou.

Sendo a hipótese nula verdadeira, as proporções populacionais são iguais. Denotando seu valor comum por p, isto é $p_1 = p_2 = p$, podemos obter um estimador para p através da ponderação dos estimadores não viciados $\widehat{p}_1$ e $\widehat{p}_2$. Dessa forma, obtemos

$$\widehat{p}_p = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}\,.$$

Substituindo os valores de p_1 e p_2 por $\widehat{p}_p$ na expressão da $Var(\widehat{p}_1 - \widehat{p}_2)$, podemos escrever, sob H_o,

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}_p(1-\widehat{p}_p)(1/n_1 + 1/n_2)}} \sim N(0,1).$$

Para concluir o teste, calculamos a quantidade z_{obs}, substituindo $\widehat{p}_1$ e $\widehat{p}_2$ por suas correspondentes estimativas. Verificamos se z_{obs} pertence à região crítica, que no caso bilateral é dada por

$$RC = \{\, z \in \mathbb{R} \,|\; z < z_{c_1} \ ou \ z > z_{c_2} \}.$$

Dado um nível de significância α, os valores z_{c_1} e z_{c_2} são obtidos da tabela da distribuição Normal padrão. Como procedimento alternativo, podemos também usar o nível descritivo para decidir sobre a aceitação ou não de H_o.

Exemplo 9.10: Para o Exemplo 9.9, testaremos

$$H_o : p_1 = p_2 \ \ versus \ \ H_a : p_1 \neq p_2 \,,$$

com p_1 e p_2 representando as proporções de crianças com verminose na população dos grupos I e II, respectivamente. Pelas informações recebidas, $n_1 = 120$, $n_2 = 260$, $\widehat{p}_{1obs} = 0{,}085$ e $\widehat{p}_{2obs} = 0{,}103$. Logo, sob H_o

$$\widehat{p}_{p_{obs}} = \frac{n_1\,\widehat{p}_{1obs} + n_2\,\widehat{p}_{2obs}}{n_1 + n_2} = \frac{120 \times 0{,}085 + 260 \times 0{,}103}{120 + 260} = 0{,}097;$$

e também,

$$\begin{aligned}\widehat{p}_{p_{obs}}(1 - \widehat{p}_{p_{obs}})(1/n_1 + 1/n_2) &= 0{,}097 \times 0{,}903 \times (1/120 + 1/260) \\ &= 0{,}0011.\end{aligned}$$

Segue então que

$$\frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{0{,}0011}} \sim N(0,1).$$

Para $\alpha = 0{,}08$ os valores z_{c_1} e z_{c_2} são calculados através das expressões

$$P(\,(\widehat{p}_1 - \widehat{p}_2)/\sqrt{0{,}0011} < z_{c_1} \,|\, H_o\,) = 0{,}04\,;$$
$$P(\,(\widehat{p}_1 - \widehat{p}_2)/\sqrt{0{,}0011} > z_{c_2} \,|\, H_o\,) = 0{,}04\,.$$

Assim,

$$RC = \{ z \in \mathbb{R} \mid z < -1{,}75 \text{ ou } z > 1{,}75 \}.$$

Fazendo os cálculos, temos que $z_{obs} = -0{,}543$ não pertence à RC. Logo, aceitamos a igualdade das proporções, ao nível 8% e concluímos que a incidência de verminose nas duas faixas etárias pode ser considerada a mesma. □

Exercícios da Seção 9.2:

1. Para se avaliar o nível de tensão ocasionada por exames escolares, doze alunos foram escolhidos e sua pulsação medida antes e depois do exame.

Instante	Estudante											
da medição	1	2	3	4	5	6	7	8	9	10	11	12
Antes	87	78	85	93	76	80	82	77	91	74	76	79
Depois	83	84	79	88	75	81	74	71	78	73	76	71

Faça um teste, com nível de significância de 1%, para verificar se existe maior tensão (isto é, maior pulsação) antes da realização dos exames. Indique as suposições necessárias.

2. Sabe-se que o tempo necessário para percorrer uma determinada rota no final da tarde pode ser estudado por um modelo Normal com desvio padrão de 17 min. Foram instalados sensores para controlar o tempo de abertura dos semáforos presentes na rota e deseja-se verificar se o tempo gasto para completar o percurso diminuiu. Estudos anteriores indicam que o tempo deve continuar se comportando segundo um modelo Normal, com mesmo desvio padrão. Com os sensores desativados, 11 veículos de mesmo ano e marca, denominado Grupo Controle, tiveram o tempo gasto no percurso anotado. Em seguida, os sensores foram ativados e outros 13 veículos (Grupo Teste) percorreram a mesma rota. Os tempos observados, em minutos, foram os seguintes:

Grupo	Tempos utilizados no percurso												
Controle	38	26	20	70	16	26	38	32	45	49	32		
Teste	17	31	28	21	50	21	20	51	10	22	18	35	29

Indique se o uso dos sensores contribui para diminuir o tempo médio de percurso utilizando o nível descritivo do teste.

3. Para verificar se duas populações têm a mesma média, amostras independentes foram retiradas. Sabendo que a população I é Normal $(\mu_1, 25)$ e a população II Normal $(\mu_2, 40)$, que conclusão pode ser tirada, ao nível 2%? Os valores obtidos foram:

População	Dados							
I	12	14	15	14	13	17	14	13
II	13	17	14	13	16	17	18	16

4. As variáveis X e Y seguem a distribuição Normal com mesma variância. Deseja-se testar se, também, têm a mesma média. Doze observações de cada variável foram escolhidas e os resultados foram os seguintes:

$$\sum_{i=1}^{12} x_i = 48, \ \sum_{i=1}^{12} y_i = 56, \ \sum_{i=1}^{12} x_i^2 = 4.900, \ \sum_{i=1}^{12} y_i^2 = 5.650.$$

Qual é a conclusão ao nível de significância de 5%?

5. Para comparar as médias de duas populações Normais, amostras aleatórias foram obtidas. Sabe-se que as variâncias populacionais são diferentes, sendo seus valores desconhecidos.

Amostra I	7	9	3	8	11	5	9
Amostra II	2	7	5	15	9	16	8

O que pode ser dito a respeito das médias das populações, com $\alpha = 0{,}05$?

6. Dois medicamentos para tratamento de infecções bucais estão sendo estudados e o melhor desempenho é definido pela rapidez em eliminar a infecção. Pacientes escolhidos ao acaso receberam um dos medicamentos e tiveram a sua cura classificada em rápida ou não. Deseja-se testar, ao nível 10%, se os medicamentos são equivalentes. Os dados obtidos são apresentados a seguir.

	Amostra	Pacientes com cura rápida
Medicamento A	50	32
Medicamento B	100	48

Qual seria sua conclusão? Indique as hipóteses do teste e as suposições necessárias.

9.3 Testes para Variância

Os testes de hipóteses considerados até aqui envolveram, como parâmetros de interesse, a média ou a proporção populacionais. Na seção anterior, vimos que, para desenvolver os testes de comparação de médias, é importante considerar o que ocorre com a variância populacional. Nesta seção, estamos interessados em estudar a dispersão dos valores em uma ou mais populações, através de suas variâncias. Para isso, apresentamos alguns testes envolvendo essas quantidades. Iniciamos com o caso de uma única população, conforme motivado pelo exemplo a seguir.

Exemplo 9.11: Sabe-se que em uma região do país a altura média é de 1,68 m, com variância 0,30 m^2. Um pesquisador acredita que a alimentação rotineira em uma cidade litorânea, sendo diferente da região como um todo, contribui para que as pessoas apresentem alturas mais homogêneas, apesar de não alterar a altura média da população da cidade. Para verificar sua suspeita, ele coletou uma amostra de 31 pessoas e obteve como estimativa para a variância o valor $s^2_{obs} = 0{,}25\,\text{m}^2$. Neste caso, o pesquisador deve realizar um teste de hipóteses relacionado à variância populacional para tirar suas conclusões. □

Considere as hipóteses nula e alternativa, dadas por:

$$H_o : \sigma^2 = \sigma^2_o\,;$$
$$H_a : \sigma^2 \neq \sigma^2_o \text{ (ou } \sigma^2 > \sigma^2_o \text{ ou } \sigma^2 < \sigma^2_o)\,.$$

Aqui, σ^2 representa a variância populacional em que estamos interessados e σ^2_o é um valor numérico particular para o parâmetro. Para testar a hipótese H_o, obtemos uma amostra de tamanho n da população e consideramos a quantidade

$$V = \frac{(n-1)S^2}{\sigma^2},$$

que envolve S^2, a variância amostral. Se a população da qual a amostra foi retirada se comporta de acordo como um modelo Normal, então, sob a hipótese H_o temos que

$$V = \frac{(n-1)S^2}{\sigma^2_o} \sim \chi^2_{n-1}.$$

O estimador S^2 é utilizado no teste pois, como visto anteriormente, é um estimador não viciado para σ^2. Baseando-se em V, construímos o teste, cuja região crítica, para o teste bilateral, terá a forma

$$RC = \{v \in \mathbb{R} \mid v < v_{c_1} \text{ ou } v > v_{c_2}\},$$

com v_{c_1} e v_{c_2} determinados a partir da distribuição Qui-quadrado com $n-1$ graus de liberdade e nível de significância α. O procedimento é ilustrado no próximo exemplo.

Exemplo 9.12: Para o Exemplo 9.11, as hipóteses de interesse são

$$H_o : \sigma^2 = 0{,}30\,;$$
$$H_a : \sigma^2 < 0{,}30\,.$$

Temos que, sob H_o, $\sigma^2 = 0{,}30$. Logo,

$$v_{obs} = \frac{(n-1)s^2_{obs}}{\sigma^2_o} = \frac{30 \times 0{,}25}{0{,}30} = 25{,}00.$$

A região crítica do teste será dada por $RC = \{v \in \mathbb{R} \mid v < v_c\}$. Utilizando a tabela da distribuição de Qui-quadrado com 30 graus de liberdade e tomando $\alpha = 0{,}05$ vem

$$P(V < v_c \mid \sigma^2 = 0{,}30) = 0{,}05 \Rightarrow v_c = 18{,}49.$$

Logo, como $v_{obs} > 18{,}49$ não rejeitamos a hipótese H_o, isto é, as alturas na cidade não são mais homogêneas do que aquelas observadas na região como um todo. Ao invés de fixar α, poderíamos calcular o nível descritivo, obtendo $\alpha^* \simeq 0{,}30$. □

Outro teste de interesse é aquele que verifica se as variâncias de duas populações são iguais. O teste de comparação de variâncias é também útil como um procedimento preliminar em testes de comparação de médias, auxiliando a escolha das técnicas adequadas.

Exemplo 9.13: Um fabricante de esferas para rolamentos desenvolveu um novo método de produção, mais barato. Entretanto, ele desconfia que os novos lotes apresentam variabilidade diferente daqueles produzidos pelo método antigo (com relação ao diâmetro das esferas). Para cada método, ele selecionou aleatoriamente 15 esferas que forneceram os seguintes diâmetros (em mm):

Método			
X (antigo)		Y (novo)	
29,9	30,1	29,8	30,3
29,8	29,9	29,8	29,9
29,8	30,0	30,4	29,7
29,7	30,0	29,8	30,3
29,9	29,6	30,5	30,4
29,8	30,4	29,6	29,1
29,9	29,9	29,3	30,0
29,9		29,4	

Algumas medidas descritivas foram calculadas, sendo $\overline{x}_{obs} = 29{,}93$ mm, $\overline{y}_{obs} = 29{,}89$ mm, $s^2_{X_{obs}} = 0{,}03$ mm^2 e $s^2_{Y_{obs}} = 0{,}19$ mm^2. Na figura que segue, apresentamos os respectivos gráficos *box-plot*, que sugerem variabilidades diferentes entre os dois métodos.

Entretanto, para podermos tirar uma conclusão objetiva, precisamos testar a hipótese de igualdade de variâncias. □

Vamos construir agora o teste de igualdade de variâncias de duas populações, representadas por X e Y, tais que $X \sim N(\mu_X, \sigma^2_X)$ e $Y \sim N(\mu_Y, \sigma^2_Y)$. Desejamos testar as hipóteses

$$H_o : \sigma^2_X = \sigma^2_Y\,;$$
$$H_a : \sigma^2_X \neq \sigma^2_Y\,.$$

Utilizaremos a quantidade

$$F = S^2_X / S^2_Y,$$

baseada nas amostras $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$, obtidas das populações de interesse, cujas variâncias estão sendo comparadas. Sob a hipótese H_o, pode ser mostrado que F segue o modelo de Fisher-Snedecor, que é caracterizado pelos graus de liberdade associados às quantidades presentes no numerador e no denominador de F, no caso, $n_1 - 1$ e $n_2 - 1$, respectivamente. Para a distribuição de Fisher-Snedecor, utilizaremos a notação $F(a, b)$, sendo a e b os graus de liberdade.

Figura 9.2: Distribuição de Fisher- Snedecor.

Probabilidades baseadas na distribuição de Fisher-Snedecor têm de ser calculadas computacionalmente e são obtidas em planilhas eletrônicas e

programas estatísticos. Para valores selecionados de n_1 e n_2, tabelas podem ser consultadas e, em geral, são construídas de forma a fornecer, para uma dada probabilidade, o valor f_c conforme mostrado na Figura 9.2. No Apêndice A, apresentamos tabelas dessa distribuição para probabilidades iguais a 0,05 e 0,95.

Assim, para um nível de significância α pré-fixado, podemos obter os valores f_1 e f_2 tais que

$$P(F < f_1 \; ou \; F > f_2) = \alpha\,,$$

com

$$F \sim F(n_1 - 1;\, n_2 - 1)\,.$$

A região crítica para o teste bilateral é dada por

$$RC = \{f \in \mathbb{R} : f < f_1 \; ou \; f > f_2\}.$$

Portanto, se $f_{obs} \in RC$, rejeitamos a hipótese de igualdade das variâncias.

Exemplo 9.14: Voltando ao exemplo anterior, queremos testar se as variâncias do diâmetro das esferas produzidas pelos métodos antigo (X) e novo (Y) são iguais ou não. Isto é,

$$H_o : \sigma_X^2 = \sigma_Y^2\,;$$
$$H_a : \sigma_X^2 \neq \sigma_Y^2\,.$$

Note que as hipóteses podem, de forma equivalente, ser expressas como

$$H_o : \frac{\sigma_X^2}{\sigma_Y^2} = 1\,;$$

$$H_a : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1\,.$$

Sob a hipótese H_o, temos que

$$F = \frac{S_X^2}{S_Y^2} \sim F(14, 14).$$

Logo, fixando $\alpha = 0{,}10$ determinamos a região crítica do teste, de modo que $P(F < f_1) = 0{,}05$ e $P(F > f_2) = 0{,}05$. Dada a forma das tabelas de Fisher-Snedecor apresentadas no Apêndice A, precisamos determinar f_1 e f_2 tais que

$$P(F > f_1) = 1 - P(F \leq f_1) = 1 - 0{,}05 = 0{,}95\,;$$
$$P(F > f_2) = 0{,}05\,.$$

Essas quantidades estão representadas nas figuras a seguir.

Da tabela da distribuição de Fisher-Snedecor, com 14 graus de liberdade para o numerador e 14 graus de liberdade para o denominador, obtemos que $f_1 = 0{,}403$ e $f_2 = 2{,}484$. Logo,

$$RC = \{f \in \mathbb{R}^+ : f < 0{,}403 \; ou \; f > 2{,}484\}.$$

Para os dados disponíveis, temos que

$$f_{obs} = s^2_{X_{obs}}/s^2_{Y_{obs}} = 0{,}03/0{,}19 = 0{,}158 \in RC.$$

Portanto, confirmando as evidências fornecidas pela análise descritiva, concluímos ao nível $\alpha = 10\%$ que existem diferenças em termos da homogeneidade dos diâmetros das esferas, dependendo do método utilizado. □

Uma peculiaridade aparece no caso de testes unilaterais, uma vez que a forma da região crítica depende de qual quantidade é considerada no numerador da expressão de F. Para esses casos, a representação das hipóteses de interesse em termos de frações evita possíveis confusões na construção de F. O exemplo, a seguir, ilustra esse procedimento.

Exemplo 9.15: Um fabricante de panetones costuma vender produtos de segunda qualidade (no que diz respeito ao formato) a preços reduzidos. Para panetones de 500 gramas, suspeita-se que o produto de segunda qualidade apresente maior variabilidade no que se refere ao peso. Para tanto, 26 panetones de primeira qualidade e 20 de segunda tiveram seus pesos aferidos. Denotaremos esses pesos

por $X_1, \ldots, X_{26}$ e $Y_1, \ldots, Y_{20}$, respectivamente. Foram calculados os valores das variâncias amostrais, sendo dados por $s^2_{X_{obs}} = 0{,}29$ e $s^2_{Y_{obs}} = 0{,}73$.

As hipóteses de interesse são:

$$H_o : \sigma^2_X = \sigma^2_Y \ \textit{versus} \ H_a : \sigma^2_X < \sigma^2_Y .$$

Para determinar a região crítica e a quantidade F corretamente, reescrevemos as hipóteses como

$$H_o : \sigma^2_X/\sigma^2_Y = 1 \ \textit{versus} \ H_a : \sigma^2_X/\sigma^2_Y < 1 .$$

A construção de F deve considerar a escolha da razão de variâncias nas hipóteses, no caso com as quantidades relacionadas a X no numerador, isto é,

$$F = S^2_X/S^2_Y ,$$

e a região crítica será da forma

$$RC = \{f \in \mathbb{R} : f < f_c\}.$$

Sob a hipótese nula, $F \sim F(25, 19)$, e para $\alpha = 0{,}05$ obtemos, da tabela da distribuição Fisher-Snedecor, $f_c = 0{,}495$.

Como

$$f_{obs} = s^2_{X_{obs}}/s^2_{Y_{obs}} = 0{,}29/0{,}73 = 0{,}356 ;$$

temos que $f_{obs} \in RC$ e, portanto, concluímos que os panetones classificados como de segunda qualidade apresentam pesos com maior variabilidade do que os panetones de primeira qualidade. □

Exercícios da Seção 9.3:

1. Supondo $X \sim F(a, b)$, encontre x_c tal que:

a. $P(X \geq x_c) = 0{,}05$ com $a = 18, b = 3$.
b. $P(X > x_c) = 0{,}05$ com $a = 3, b = 18$.
c. $P(X > x_c) = 0{,}05$ com $a = 180, b = 192$.
d. $P(X > x_c) = 0{,}95$ com $a = 5, b = 12$.
e. $P(X \geq x_c) = 0{,}95$ com $a = 30, b = 40$.

2. Uma linha de montagem produz peças cujos pesos, em gramas, obedecem ao modelo Normal com variância 30 g^2. Os equipamentos foram modernizados e, para verificar se o processo continua sob controle, foi tomada uma amostra de 23 peças, que forneceu $s^2_{obs} = 40$ g^2. Existem evidências indicando que a variância mudou, considerando $\alpha = 10\%$?

3. Uma panificadora produz determinado tipo de pão, cujo peso médio é de 190 gramas, com desvio padrão de 18 gramas. Devido a mudanças na política cambial, que ocasionou aumento no preço do trigo, alguns ingredientes da receita foram substituídos. Uma equipe do governo resolveu verificar se a variabilidade no peso do produto aumentou e escolheu, aleatoriamente, 16 unidades, medindo o peso de cada uma. O peso médio obtido da amostra foi de 102 gramas e o desvio padrão foi de 24,5 gramas. Qual a conclusão para $\alpha = 10\,\%$.

4. Para comparar o grau de diversidade de duas populações primitivas, uma medida antropométrica foi obtida em fósseis coletados em sítios arqueológicos, fornecendo a tabela a seguir.

Característica	Sítio A ($n = 17$)	Sítio B ($n = 23$)
Média (cm)	15,12	12,21
Variância (cm^2)	0,124	0,184

O que pode ser concluído a respeito das variâncias? E das médias populacionais?

9.4 Análise de Variância

Consideramos nesta seção o caso de comparação de três ou mais populações, definidas por uma variável qualitativa (fator) através de testes com as correspondentes médias. Não abordaremos a situação com dois ou mais fatores neste texto e o leitor interessado poderá consultar as referências mencionadas na bibliografia. Iniciamos com o caso em que as amostras de cada população têm o mesmo tamanho.

Exemplo 9.16: A gerência de um depósito que armazena cargas aéreas de pequeno porte está estudando o peso das cargas que chegam ao seu terminal no interior de São Paulo. Usualmente, o terminal recebe 4 tipos de cargas: doméstica (D), administrativa (A), equipamentos industriais (E) e outros tipos (O). Deseja-se verificar se, em média, existem diferenças entre os pesos dos 4 tipos de cargas. Ao longo de 1 mês, cargas foram colhidas aleatoriamente e seus pesos foram aferidos, fornecendo os dados (em kg):

Tipo de Carga			
D	A	E	O
24,9	27,9	38,4	23,8
20,4	28,1	38,6	25,3
24,2	28,4	41,2	23,5
22,3	25,3	43,9	27,6
20,3	29,3	40,2	25,5
24,0	28,5	40,2	23,9
23,5	27,9	37,3	22,6

Estamos interessados em comparar os quatro grupos com diferentes tipos de carga. Descritivamente, observamos que os pesos de cargas do tipo E apresentam os maiores valores dentre os quatro grupos. Algumas medidas descritivas para os dados estão apresentadas na tabela a seguir.

Medidas Descritivas	Tipo de Carga			
	D	A	E	O
Média	22,8	27,9	40,1	24,6
Mediana	23,5	28,1	40,2	23,9
Desvio-Padrão	1,9	1,2	2,2	1,6
Mínimo	20,3	25,3	37,31	22,6
Máximo	24,9	29,3	43,9	27,6

Os valores descritivos também sugerem que cargas do grupo E tendem a ter maior peso. Note que o menor peso observado para esse tipo de carga é maior que os máximos observados para todos os outros tipos. Após a carga do tipo E, as cargas do tipo A parecem ter, em média, maior peso. A carga de tipo D apresenta, em média, o menor peso. Salientamos que os desvios padrões amostrais encontram-se razoavelmente próximos uns dos outros. □

Para estudarmos a situação apresentada no exemplo anterior, consideramos um modelo estatístico, em que cada observação Y pode ser decomposta em duas componentes: *sistemática* e *aleatória*, esta última representando variações individuais e todos os fatores que não são explicados pela parte sistemática. Matematicamente, podemos escrever

$$Y = \mu + e\,.$$

Assim, se Y representa a observação associada a uma unidade experimental, a parte sistemática μ pode ser vista como a média populacional, que é fixa, e a parte aleatória e como a informação referente a outros fatores que podem influir nas observações mas não são incorporadas em μ.

Suponha que estamos interessados em comparar as médias de K populações, isto é, queremos testar

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_K\,;$$
$$H_a : \text{pelo menos uma das médias } \mu_i \text{ é diferente das demais.}$$

Para tanto, obtemos K amostras independentes, com m indivíduos em cada uma delas. Nesta situação, o modelo estatístico para a j-ésima unidade experimental, da i-ésima população é dado por

$$\text{Modelo 1: } Y_{ij} = \mu_i + e_{ij}\,,\quad i = 1, \ldots, K;\ j = 1, \ldots, m.$$

Caso a hipótese H_o seja verdadeira, teremos que todas as médias para as K populações serão iguais a um valor comum μ. O modelo pode, então, ser escrito como

$$\text{Modelo 0: } Y_{ij} = \mu + e^*_{ij},\quad i = 1, \ldots, K;\ j = 1, \ldots, m.$$

Uma forma de levar em conta a informação não explicada pela forma sistemática é através das somas de quadrado

$$\sum_{i=1}^{K}\sum_{j=1}^{m} e_{ij}^2 = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \mu_i)^2 \qquad \text{e} \qquad \sum_{i=1}^{K}\sum_{j=1}^{m}(e^*_{ij})^2 = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \mu)^2.$$

Essas expressões envolvem as quantidades desconhecidas μ_i, $i = 1, \ldots, K$ e μ. Utilizaremos os dados para obter as estimativas correspondentes. Levando-se em conta que, no Modelo 1, estamos supondo diferentes médias para as K populações, consideramos os dados oriundos de cada uma dessas populações para estimar a correspondente média. Assim, segundo o Modelo 1 temos

$$\widehat{\mu}_i = \frac{1}{m}\sum_{j=1}^{m} Y_{ij} = \overline{Y}_i\,,\ i = 1, \ldots, K.$$

Para o Modelo 0, como assumimos que todas as populações têm a mesma média, utilizamos o estimador

$$\widehat{\mu} = \frac{1}{mK}\sum_{i=1}^{K}\sum_{j=1}^{m} Y_{ij} = \overline{Y}.$$

Substituindo as estimativas acima nas correspondentes somas de quadrados apresentadas anteriormente, definimos as quantidades *soma de quadrados dentro* (SQD) e *soma de quadrados total* (SQT), da seguinte forma:

$$\text{SQD} = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \widehat{\mu}_i)^2 = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \overline{Y}_i)^2 = \sum_{i=1}^{K}\sum_{j=1}^{m} Y_{ij}^2 - m\sum_{i=1}^{K}\overline{Y}_i^2\,;$$

$$\text{SQT} = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \widehat{\mu})^2 = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{m} Y_{ij}^2 - mK\overline{Y}^2.$$

É importante ressaltar que as expressões resultantes fornecem uma maneira mais conveniente de se calcular, via computador ou manualmente, cada uma dessas somas de quadrados.

A diferença entre SQT e SQD representa a *soma de quadrados entre* e será denotada por SQE, isto é,

$$\text{SQE} = \text{SQT} - \text{SQD}.$$

Das expressões para a soma de quadrados total e de dentro, segue que:

$$\text{SQE} = m\sum_{i=1}^{K}(\overline{Y}_i - \overline{Y})^2 = m(\sum_{i=1}^{K}\overline{Y}_i^2 - K\overline{Y}^2).$$

Cada uma das somas de quadrados envolve um certo número de quantidades que estão sendo estimadas. Por exemplo, SQT contém $\overline{Y}$ e SQD contém $\overline{Y}_i$, $i = 1, \ldots, K$. Levando este fato em consideração e o número de observações nas amostras, definimos os correspondentes quadrados médios:

$$\text{QMT} = \frac{\text{SQT}}{Km-1}\text{: quadrado médio total;}$$

$$\text{QMD} = \frac{\text{SQD}}{Km-K} = \frac{\text{SQD}}{K(m-1)}\text{: quadrado médio dentro;}$$

$$\text{QME} = \frac{\text{SQE}}{K-1}\text{: quadrado médio entre.}$$

Note que, nesse caso, é preciso calcular as três quantidades anteriores pois QMT não é igual à soma de QMD com QME.

O teste estatístico para a hipótese H_o envolve os quadrados médios. Se a hipótese H_o não for verdadeira, então, o Modelo 1 deve ser mais adequado aos dados do que o Modelo 0. Em outras palavras, os resíduos produzidos pelo Modelo 1 serão menores que os do Modelo 0. Analisando sob este ângulo, podemos interpretar QME como sendo, em termos quantitativos, a informação contida nos dados que é captada pelo Modelo 1, enquanto que o QMD representa a informação não explicada pelo Modelo 1. Portanto, se QME for grande comparado à QMD, a parte sistemática do Modelo 1 estará captando grande parte da informação dos dados e a hipótese H_o deverá ser rejeitada. Definimos, então, a quantidade

$$F = \frac{\text{QME}}{\text{QMD}}\,.$$

Quanto maior for o valor de F, maior será QME comparado a QMD e, assim, maiores as evidências contra H_o. Para caracterizarmos o valor crítico a partir do qual rejeitamos H_o, precisamos encontrar a distribuição de probabilidade para F. Supondo as seguintes condições:

- Y_{ij} são variáveis aleatórias independentes;
- Todas as K populações têm variâncias iguais a σ^2;
- $Y_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, K$ e $j = 1, \ldots, m$,

pode ser mostrado que

$$F \sim F(K-1,\, K(m-1))$$

isto é, a quantidade F tem distribuição de Fisher-Snedecor com $K-1$ e $K(m-1)$ graus de liberdade. Temos, agora, condições de encontrar o valor crítico f_c e determinar a região crítica do teste, que será da forma

$$RC = \{f \in \mathbb{R}^+ : f > f_c\}.$$

Das três suposições feitas, a mais importante é a segunda, $Var(Y_{ij}) = \sigma^2$, para $i = 1, \ldots, K$ e $j = 1, \ldots, m$, que tem o nome técnico de *homocedasticidade*. A suposição de normalidade é importante em termos teóricos, mas, muitas vezes, na prática, o teste pode ainda ser utilizado quando ela não for válida, principalmente, se as amostras forem grandes. Nesses casos, o Teorema Central do Limite pode ser utilizado para justificar o uso da distribuição de Fisher-Snedecor. Caso a suposição de homocedasticidade não seja verdadeira, técnicas

alternativas podem ser utilizadas. Algumas delas envolvem aplicar uma transformação logarítmica ou quadrática aos dados. Esse assunto envolve técnicas mais avançadas e não será abordado nesse livro.

A discussão sobre o comportamento dos erros e das somas de quadrados é resumida na Tabela 9.2 a seguir.

Tabela 9.2: Tabela de Análise de Variância (ANOVA).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Entre	$K-1$	SQE	QME	QME/QMD
Dentro	$K(m-1)$	SQD	QMD	
Total	$Km-1$	SQT		

A tabela ANOVA fornece como subproduto um estimador para a variância populacional σ^2, baseado na suposição de homocedasticidade. Nesse caso, a variância amostral para o i-ésimo grupo,

$$S_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_{ij} - \overline{Y}_i)^2,$$

pode ser usada para construir um estimador da variância populacional. Isto é feito combinando esses valores através da média ponderada de $S_1^2, \ldots, S_K^2$:

$$S_c^2 = \frac{(m-1)S_1^2 + \cdots + (m-1)S_K^2}{(m-1) + \cdots + (m-1)} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \overline{Y}_i)^2}{K(m-1)}.$$

A expressão obtida para S_c^2 é a mesma que encontramos para QMD. Note ainda que a expressão de QMT também é um estimador para σ^2, uma vez que

$$\text{QMT} = \frac{\text{SQT}}{Km-1} = \frac{1}{Km-1}\sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij} - \overline{Y})^2 = S^2.$$

Ou seja, QMT nada mais é do que a variância amostral S^2 para uma amostra composta pelo conjunto de todas as observações dos K grupos combinados.

Exemplo 9.17: Para os dados apresentados no Exemplo 9.16, temos $K = 4$ grupos e $m = 7$ observações por grupo. Além disso, obtemos $\overline{Y}_1 = 22{,}8$; $\overline{Y}_2 = 27{,}9$; $\overline{Y}_3 = 40{,}1$ e $\overline{Y}_4 = 24{,}6$. A média geral é $\overline{Y} = 28{,}9$. Cálculos intermediários podem ser, facilmente, feitos em uma planilha eletrônica ou calculadora fornecendo:

$$\sum_{i=1}^{4}\sum_{j=1}^{7} Y_{ij}^2 = 24.672{,}42 \quad \text{e} \quad \sum_{i=1}^{4}\overline{Y}_i^2 = 3.513{,}80.$$

Usando as fórmulas de cálculo apresentadas anteriormente, obtemos

$$\text{SQD} = \sum_{i=1}^{K}\sum_{j=1}^{m} Y_{ij}^2 - m\sum_{i=1}^{K}\overline{Y}_i^2 = 24.672{,}42 - 7 \times 3.513{,}80 = 75{,}82;$$

$$\text{SQE} = m\left(\sum_{i=1}^{K}\overline{Y}_i^2 - K\overline{Y}^2\right) = 7 \times (3.513{,}80 - 4 \times 28{,}86^2) = 1.275{,}41;$$

$$\text{SQT} = \sum_{i=1}^{K}\sum_{j=1}^{m} Y_{ij}^2 - mK\overline{Y}^2 = 24.672{,}42 - 7 \times 4 \times 28{,}86^2 = 1.351{,}23.$$

Uma vez calculadas duas das somas de quadrados acima, obtemos, sem dificuldade, a terceira. A tabela ANOVA é apresentada a seguir.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Entre	3	1.275,41	$\frac{1.275{,}41}{3} = 452{,}14$	$\frac{425{,}4}{3{,}16} = 134{,}54$
Dentro	24	75,82	$\frac{75{,}82}{24} = 3{,}16$	
Total	27	1.351,23		

Através da distribuição de Fisher-Snedecor, com 3 e 24 graus de liberdade e, considerando $\alpha = 5\%$, obtemos $f_c = 3{,}009$. Logo, como calculamos $f_{obs} = 134{,}54 > f_c$, concluímos que, ao nível de significância de 5%, as médias de peso dos grupos são diferentes, confirmando as observações descritivas feitas anteriormente. □

Grupos de tamanhos diferentes

No desenvolvimento anterior, supomos que os K grupos têm todos o mesmo tamanho. Podemos considerar uma situação mais geral em que isto não acontece. Vamos denotar por n_i o número de elementos do grupo i. Neste caso, o total de indivíduos nos K grupos será igual a

$$n = n_1 + \cdots + n_K.$$

Todos os resultados anteriores permanecem válidos, mas modificações algébricas são necessárias nas expressões que agora serão escritas da seguinte forma:

$$\text{SQD} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{K} n_i \overline{Y}_i^2;$$

$$\text{SQE} = \sum_{i=1}^{K} n_i(\overline{Y}_i - \overline{Y})^2 = \sum_{i=1}^{K} n_i \overline{Y}_i^2 - n\overline{Y}^2\,;$$

$$\text{SQT} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i} Y_{ij}^2 - n\overline{Y}^2.$$

Note que, nesse caso, as médias geral e dos grupos são dadas por:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} Y_{ij}\,;$$

$$\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\,,\ i = 1, \ldots, K.$$

A Tabela de Análise de Variância sofre poucas modificações, sendo dada por

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Entre	$K-1$	SQE	QME	QME/QMD
Dentro	$n-K$	SQD	QMD	
Total	$n-1$	SQT		

com $F \sim F(K-1, n-K)$.

Exemplo 9.18: O volume de vendas, no ramo de vestuário, tem se mantido estável de ano para ano, mas acredita-se que sofra mudança de um quadrimestre para outro, dentro de um mesmo ano. Através de uma metodologia adequada, foi criado um índice que reflete a quantidade vendida. Em cada um dos quadrimestres do ano, foram escolhidas aleatoriamente algumas empresas de mesmo porte e seus índices de venda foram calculados (ver abaixo).

Quad1	Quad2	Quada3
114,7	144,7	153,1
144,7	173,4	192,5
119,1	154,2	145,5
113,7	154,7	168,8
108,9	125,9	141,5
96,7	119,5	141,2
87,6	155,7	189,6
132,4	213,9	178,4
	156,2	208,6
	159,0	

O comportamento das vendas pode ser visualizado na próxima figura.

Uma rápida avaliação dos *box-plot* mostra o primeiro quadrimestre com os menores índices. Os outros dois quadrimestres apresentam valores um pouco mais próximos.

O modelo de análise de variância pode ser aplicado para verificar a significância estatística das diferenças observadas para os dados. Aqui, temos $K = 3$, $n_1 = 8$, $n_2 = 10$ e $n_3 = 9$. Fazendo os cálculos preliminares, obtemos que $\overline{Y}_1 = 114{,}7$; $\overline{Y}_2 = 155{,}7$; $\overline{Y}_3 = 168{,}8$ e $\overline{Y} = 147{,}9$. Com os dados disponíveis e a ajuda de uma planilha eletrônica obtemos

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i} Y_{ij}^2 = 617.359{,}68 \quad \text{e} \quad \sum_{i=1}^{3} n_i \overline{Y}_i^2 = 604.207{,}68\,.$$

Então,

$$\text{SQD} = \sum_{i=1}^{3}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{3} n_i \overline{Y}_i^2 = 617.359{,}68 - 604.207{,}68 = 13.332{,}00\,;$$

$$\text{SQE} = \sum_{i=1}^{K} n_i \overline{Y}_i^2 - n\overline{Y}^2 = 604.207{,}68 - 27 \times 147{,}93^2 = 13.344{,}90\,;$$

e, com relação à variação total,

$$\text{SQT} = 13.332{,}00 + 13.344{,}90 = 26.676{,}90\,.$$

Com esses valores calculados, construímos a tabela ANOVA:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Entre	2	13.344,90	$\frac{13.344{,}90}{2} = 6.672{,}45$	$\frac{6.672{,}45}{555{,}50} = 12{,}01$
Dentro	24	13.332,00	$\frac{13.332{,}00}{24} = 555{,}50$	
Total	26	26.676,90		

O teste fornece o valor $f_{obs} = 12{,}01$ que deve ser comparado com o valor crítico obtido de uma distribuição Fisher-Snedecor com 2 e 24 graus de liberdade. Considerando $\alpha = 5\%$, obtemos $f_c = 3{,}403$. Tendo em vista que $f_{obs} > f_c$, concluímos que existe diferença nas médias de venda dos quadrimestres. □

Exercícios da Seção 9.4:

1. Três diferentes bancos possuem agências de mesmo porte em uma avenida movimentada de Salvador, BA. Para testar se essas agências têm movimento médio equivalente, foi escolhida uma semana típica de trabalho e o desempenho, nesses dias, foi registrado. Os dados obtidos, em milhares de reais, estão apresentados na tabela a seguir:

Banco		
1	2	3
146,4	194,3	173,7
199,2	227,2	246,5
179,5	203,4	289,8
98,4	111,8	127,4
263,7	275,0	265,6

Qual seria a sua conclusão ao nível $\alpha = 5\%$?

2. Uma agência de empregos deseja verificar o grau de satisfação de seus clientes. Para tanto, escolheu aleatoriamente domicílios de famílias de bairros classe A, B e C, que fizeram uso da agência e solicitou que um questionário fosse preenchido pela pessoa responsável na família. Os questionários foram devidamente codificados, a fim de fornecer um índice de satisfação que varia de 1 a 5 (totalmente satisfeito). Os resultados estão apresentados a seguir. Qual seria sua conclusão, considerando $\alpha = 0{,}05$?

Classe					
A		B		C	
2,7	3,7	4,0	4,1	1,5	2,8
4,3	4,6	2,3	4,2	1,7	3,1
3,4	4,7	2,5	3,5	2,7	2,4
2,9	3,5	2,5	4,2	2,5	
4,5	3,8				

3. A fim de verificar o efeito de quatro tipos de propaganda de uma determinada marca de goma de mascar, crianças foram atribuídas aleatoriamente a cada uma de 4 salas que mostravam desenhos animados, com intervalos regulares em que as correspondentes propagandas eram inseridas. Após a sessão, as

crianças foram entrevistadas por psicólogos, que atribuíram um índice de assimilação a cada criança. Quanto maior esse índice, maior seria a lembrança do produto. Os dados são apresentados a seguir.

	Tipo de Propaganda			
	I	II	III	IV
	15 7	22 22	8 15	17 10
	7 6	21 16	8 11	16 11
	6 7	23 15	7 16	19 18
	10 6	20 22	10 8	11 11
	5 6	18 18	13 15	11 10
	5 8	21 22	8 8	13 19
$\sum x_i$	88	240	127	166
$\sum x_i^2$	730	4.876	1.465	2.444

9.5 Regressão Linear Simples

No Capítulo 5, definimos o coeficiente de correlação como uma medida de dependência linear entre duas variáveis. Em muitas situações, além de estarmos interessados em saber se existe relação entre duas variáveis, podemos desejar estabelecer uma relação de causalidade. Isto é, queremos quantificar qual é a mudança observada em uma das variáveis quando variamos os valores da outra.

Exemplo 9.19: Em uma dada região de Bocaina-SP, acredita-se que o gado alimentado em um determinado pasto tem um ganho de peso maior que o usual. Estudos de laboratório detectaram uma substância no pasto e deseja-se verificar se ela pode ser utilizada para melhorar o ganho de peso dos bovinos. Foram escolhidos 15 bois de mesma raça e idade, e cada animal recebeu uma determinada concentração da substância X (em mg/l). O ganho de peso após 30 dias, denotado por Y, foi anotado e os dados foram os seguintes (em kg):

X	0,2	0,5	0,6	0,7	1,0	1,5	2,0
Y	9,4	11,4	12,3	10,2	11,9	13,6	14,2

X	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0
Y	16,2	16,2	17,7	18,8	19,9	22,5	24,7	23,1

Observando a tabela de dados, notamos que, de fato, à medida que aumenta a concentração da substância, ocorre um aumento no ganho de peso.

Conforme apresentado no Capítulo 5, o coeficiente de correlação linear entre a concentração X e o ganho de peso Y é calculado por:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\,\overline{x}_{obs}\,\overline{y}_{obs}}{\sqrt{\left[\sum_{j=1}^{n} x_j^2 - n\,\overline{x}_{obs}^2\right]\left[\sum_{j=1}^{n} y_j^2 - n\,\overline{y}_{obs}^2\right]}}$$

$$= \frac{785{,}55 - 15 \times 2{,}70 \times 16{,}14}{\sqrt{(163{,}39 - 15 \times 2{,}70^2)(4.239{,}43 - 15 \times 16{,}14^2)}} = 0{,}99.$$

Desta forma, vemos que a variação do peso é sensivelmente influenciada pela variação da concentração da substância.

Para observarmos como as variáveis se relacionam, construímos um gráfico de dispersão, apresentado na figura a seguir.

Nota-se que os pontos tendem a se alinhar sobre uma reta, cuja inclinação reflete o sinal positivo observado no coeficiente de correlação calculado. □

Utilizando o desenvolvimento da Análise de Variância apresentado na seção anterior, consideramos para o conjunto de valores (X_i, Y_i), $i = 1, \ldots, n$, o seguinte modelo estatístico:

$$Y_i = g(X_i) + e_i\,.$$

Isto é, o comportamento de Y_i é explicado em parte por X_i, através da função $g(X_i)$ e, em uma outra parte não captada por essa função, representada por e_i. Várias opções para $g(X_i)$ podem ser utilizadas, mas a que define o modelo de regressão linear simples é

$$g(X_i) = \alpha + \beta X_i.$$

Portanto, dado um valor fixado x_i para X_i, o modelo pode ser reescrito como

$$\boxed{Y_i = \alpha + \beta x_i + e_i.}$$

Além disso, supomos que os termos e_i, $i = 1, \ldots, n$, são independentes e distribuídos conforme um modelo Normal de média 0 e variância σ^2. Dessa forma, fixado $X_i = x_i$, as variáveis Y_i são independentes e

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, \ldots, n.$$

Em modelos de regressão, a variável Y_i é, comumente, denominada *variável resposta* ou *variável dependente*, ao passo que a variável X_i é chamada *variável independente*, *explicativa* ou ainda *covariável*. Os parâmetros do modelo em que temos maior interesse são α e β. Eles têm interpretações muito úteis na prática. O parâmetro α é o valor esperado para a variável dependente Y_i quando X_i é igual a zero. Para a interpretação do parâmetro β considere dois valores para X_i, dados por x e $x+1$, e represente por $E(Y \mid X = x)$ o valor esperado da variável resposta quando $X = x$. Então,

$$\begin{aligned} E(Y_i \mid X_i = x+1) &= \alpha + \beta(x+1) \\ &= (\alpha + \beta x) + \beta \\ &= E(Y_i \mid X_i = x) + \beta\,. \end{aligned}$$

Logo,

$$\beta = E(Y_i \mid X = x+1) - E(Y_i \mid X = x)\,,$$

indicando que β representa o acréscimo *esperado* na variável resposta, quando a covariável é acrescida de uma unidade. Isto nos fornece uma idéia a respeito da intensidade com a qual a covariável atua na resposta.

A estimação de α e β pode ser feita através do *método de mínimos quadrados*, que consiste em minimizar a soma dos quadrados dos resíduos obtidos através da diferença entre valores observados y e valores esperados $E(Y \mid X = x)$, calculados para cada $X = x$. A Figura 9.3 ilustra essas quantidades. Note que, caso o ajuste fosse perfeito, todos os pontos estariam alinhados sobre a reta e os resíduos e seriam todos iguais a zero.

Figura 9.3: Resíduos no modelo de regressão linear simples.

Como em geral os pontos não estão perfeitamente alinhados, escolhemos a "melhor" reta possível no sentido de minimizar a soma de quadrados $SQ(\alpha, \beta)$, dada por

$$\begin{aligned} SQ(\alpha, \beta) &= \sum_{i=1}^{n} (y_i - E(Y_i \mid X_i = x_i))^2 \\ &= \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} e_i^2. \end{aligned}$$

Matematicamente, temos que resolver o sistema de equações envolvendo as derivadas de $SQ(\alpha, \beta)$ em relação a α e β. Deixamos as contas a cargo do leitor, apresentando aqui a solução para o sistema que fornecerá os estimadores de mínimos quadrados para α e β. Temos

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}\,;$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

Exemplo 9.20: Vimos, no Exemplo 9.19, que o diagrama de dispersão sugere que uma reta pode ser utilizada para representar o efeito da concentração de uma certa substância no pasto (X), no ganho de peso bovinos (Y).

Para obtermos essa reta, precisamos calcular as estimativas de $\widehat{\alpha}$ e $\widehat{\beta}$. Dos dados fornecidos obtemos:

$$n = 15,\ \sum_{i=1}^{n} x_i y_i = 785{,}55;\ \sum_{i=1}^{n} x_i^2 = 163{,}39;\ \overline{x} = 2{,}70 \text{ e } \overline{y} = 16{,}14.$$

Logo,

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{785{,}55 - 15 \times 2{,}70 \times 16{,}14}{163{,}39 - 15 \times 2{,}70^2} = 2{,}44\,;$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 16{,}14 - 2{,}44 \times 2{,}70 = 9{,}55.$$

Portanto, dado $X = x_i$, a reta ajustada fornece valores $\widehat{y}_i$ dados por

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}\,x_i = 9{,}55 + 2{,}44\,x_i.$$

A figura a seguir mostra os dados originais e a reta ajustada. O gráfico sugere que o modelo de regressão linear simples apresenta um ajuste adequado aos dados.

A interpretação dos valores estimados é feita da seguinte forma. O ganho de peso esperado em bovinos que não recebem a substância X é 9,55 kg (obtido substituindo $x_i = 0$ na equação calculada acima). Por outro lado, um aumento de 1 mg/l na concentração de X implica em um ganho médio esperado de 2,44 kg.

Testes de hipóteses, envolvendo os parâmetros do modelo de regressão linear simples, baseiam-se na decomposição da variação total, discutida na seção anterior.

O principal teste de interesse é verificar se a covariável influencia na resposta, o que é equivalente a testar

$$H_o : \beta = 0 \textit{ versus } H_a : \beta \neq 0.$$

Caso H_o não seja rejeitada, adotamos o modelo

$$\text{Modelo 0: } Y_i = \alpha + e_i, \qquad i = 1, \ldots, n\,;$$

caso H_o seja rejeitada, o modelo é

$$\text{Modelo 1: } Y_i = \alpha + \beta\,X_i + e_i, \qquad i = 1, \ldots, n.$$

Através do Modelo 0 obtemos a soma de quadrados total, dada por

$$\text{SQT} = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

que contém a variação total contida nos dados. Por outro lado, o Modelo 1 gera a *soma de quadrados residual*

$$\text{SQRes} = \sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}\, x_i)^2,$$

que contém a variação dos dados não explicada por esse modelo. A diferença entre as duas somas de quadrados fornece a *soma de quadrados da regressão*, dada por

$$\text{SQReg} = \text{SQT} - \text{SQRes} = \widehat{\beta}^2 \times \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Para estabelecer os graus de liberdade associados às somas de quadrados, precisamos levar em conta as estimativas envolvidas em suas expressões. Assim, SQT envolve a média $\overline{y}$, e assim, temos $n-1$ graus de liberdade associados a essa quantidade; SQRes envolve duas estimativas, $\widehat{\alpha}$ e $\widehat{\beta}$, de forma que teremos $n-2$ graus de liberdade. Para a SQReg, restam $n-1-(n-2)=1$ grau de liberdade. Consequentemente, definimos os quadrados médios por

$$\text{QMT} = \frac{\text{SQT}}{n-1} = S^2,\ \text{QMRes} = \frac{\text{SQRes}}{n-2} \quad \text{e} \quad \text{QMReg} = \frac{\text{SQReg}}{1}.$$

Seguindo os passos da seção anterior, utilizamos

$$F = \frac{\text{QMReg}}{\text{QMRes}}$$

para testar as hipóteses de interesse. Pode ser mostrado que F tem distribuição de Fisher-Snedecor com 1 e $n-2$ graus de liberdade, isto é, $F \sim F(1,\, n-2)$.

Em resumo, da mesma forma que na seção anterior, podemos apresentar as informações apresentadas em uma tabela ANOVA, específica para o modelo $Y_i = \alpha + \beta X_i + e_i$, dada por

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Regressão	1	SQReg	QMReg	QMReg/QMRes
Residual	$n-2$	SQRes	QMRes	
Total	$n-1$	SQT		

Exemplo 9.21: (Continuação dos Exemplos 9.19 e 9.20) No estudo da relação entre ganho de peso de bovinos (X) e a concentração de uma substância (Y), estabelecemos uma reta de regressão.

Para verificar a evidência estatística do modelo realizamos um teste de hipóteses:

$$H_o : \beta = 0 \quad \textit{versus} \quad H_a : \beta \neq 0.$$

Os valores de QMReg e QMRes podem ser calculados com o uso de uma planilha eletrônica, conforme a tabela seguinte

i	$(y_i - \widehat{\alpha} - \widehat{\beta}\, x_i)^2$	$(x_i - \overline{x})^2$
1	0,41	6,25
2	0,39	4,84
3	1,65	4,41
4	1,12	4,00
5	0,01	2,89
6	0,15	1,44
7	0,05	0,49
8	0,30	0,04
9	0,45	0,09
10	0,15	0,64
11	0,26	1,69
12	0,40	3,24
13	0,56	5,29
14	2,98	7,84
15	1,20	10,89
Total	10,09	54,04

Com base nos valores apresentados na última linha da tabela, temos,

$$\text{SQReg} = \widehat{\beta}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 = 2{,}44^2 \times 54{,}04 = 321{,}73\,;$$

$$\text{SQRes} = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta}\, x_i)^2 = 10{,}09\,.$$

A tabela de ANOVA para o modelo de regressão proposto é dada por:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F
Regressão	1	321,73	321,73	412,47
Residual	13	10,09	0,78	
Total	14	331,82		

Para $\alpha = 0{,}05$ obtemos, da distribuição de Fisher-Snedecor com 1 e 13 graus de liberdade, $f_c =$ 4,667. Como $f_{obs} = 412{,}47 > f_c$, rejeitamos a hipótese H_o e concluímos que existem evidências estatísticas de que a concentração da substância X, de fato, altera o ganho de peso dos bovinos. □

Exercícios da Seção 9.5:

1. Um estudo deseja avaliar o efeito de determinado treinamento no tempo de reação de atletas submetidos a um certo estímulo. O treinamento consiste na repetição de um movimento e foi utilizada uma amostra de 37 atletas. Para cada atleta foi atribuído um certo número de repetições (X) e, então, foi medido o tempo de reação (Y), em milisegundos. Uma reta de mínimos quadrados foi ajustada aos dados, fornecendo a equação

$$\hat{y}_i = 80{,}5 - 0{,}90 x_i, \qquad i = 1, \ldots, n.$$

Interprete as estimativas de α e β.

2. Para verificar o efeito da variável X sobre a variável Y, foi realizado um experimento que forneceu os pares (x_i, y_i) dados por (3; 13,3), (7; 24,3), (5;15,9), (2; 12,8), (9; 29,6), (7; 29,5), (3; 14,5), (5; 23,3), (8; 32,6), (2;12,0) e (1; 4,6). Obtenha a reta ajustada. Construa o diagrama de dispersão, baseando-se nos pares de valores fornecidos e, em seguida, desenhe a reta ajustada. Baseando-se apenas no gráfico, você diria que o ajuste é adequado?

3. Para verificar se existe relação entre a renda familiar (em salários mínimos) e o número de filhos, foi coletada uma amostra de 8 famílias em uma cidade. Os resultados obtidos estão na tabela a seguir:

Família	1	2	3	4	5	6	7	8
Renda	12	14	15	17	23	27	34	43
Filhos	3	2	2	1	1	0	0	0

a. Que conclusões podem ser tiradas, baseando-se em um diagrama de dispersão e no coeficiente de correlação?

b. Calcule a reta de mínimos quadrados e interprete os parâmetros. Verifique se a renda influi no número de filhos, utilizando $\alpha = 5\%$.

9.6 Exercícios

1. A seqüência de operações executadas por um operário para realizar uma certa tarefa está sendo estudada. Para tanto, 9 operários foram sorteados e mediu-se o tempo necessário, em minutos, para que cada um realizasse a tarefa, com os dois tipos de seqüências. Suponha que o modelo Normal é adequado.

Operário	1	2	3	4	5	6	7	8	9
Atual	24	25	27	22	23	28	26	28	29
Nova	21	23	28	27	24	26	25	22	23

Baseando-se nos dados fornecidos, você diria que houve diminuição no tempo médio para a realização da tarefa? Use $\alpha = 5\%$.

2. Para se aferir o consumo de combustível, entre duas marcas de automóveis com mesmas características, escolheu-se 8 carros de cada marca e anotou-se o consumo após 100 quilômetros percorridos em uma estrada. Os resultados estão abaixo:

Marca	Consumo (km/l)							
XWX	9,5	9,4	9,6	9,1	9,3	9,9	9,8	10,1
YWY	9,0	9,3	8,6	8,1	8,3	8,9	8,8	7,9

Fazendo as suposições necessárias, verifique se o consumo médio das duas marcas é o mesmo. Use $\alpha = 5\%$. Admita que as marcas tenham a mesma variabilidade.

3. O desempenho em duas classes de Estatística está sendo comparado através do resultado dos dez melhores alunos de cada turma.

Classe	Notas									
I	8,5	7,5	7,0	6,5	8,5	9,5	9,0	9,0	8,5	10,0
II	7,0	7,5	8,5	9,5	9,0	8,5	8,0	8,5	9,5	9,5

Admitindo que as variabilidades das classes são iguais, pode ser dito que elas têm o mesmo desempenho médio? Faça as suposições necessárias e utilize $\alpha = 2\%$.

4. O salário de recém formados em Veterinária foi amostrado em duas cidades. Na cidade A, 10 profissionais foram sorteados e na cidade B, 15. Os resultados, em salários mínimos, são apresentados a seguir.

Cidade A: 7,3; 6,6; 6,8; 7,4; 8,3; 6,5; 7,9; 8,7; 8,1; 8,5

Cidade B: 6,5; 7,8; 8,2; 6,9; 7,9; 9,7; 9,1; 9,5; 7,4; 8,0; 6,9; 7,9; 8,4; 9,3; 9,5

Admitindo que a variabilidade do salário seja a mesma nas duas cidades e que o modelo Normal é adequado, verifique se a cidade B paga, em média, melhor do que A. Obtenha o nível descritivo e tome sua decisão utilizando um nível de significância $\alpha = 5\%$.

5. Motoristas novatos e experientes participaram de um experimento para avaliar se o tempo de habilitação altera o desgaste das pastilhas de freio. Dez motoristas de cada tipo foram escolhidos e observou-se o número de quilômetros até a troca das pastilhas. Indicando por X a *resposta dos novatos* e por Y a *dos experientes*, os resultados amostrados (em milhares de km) foram os seguintes:

$$\sum_{i=1}^{10} x_i = 98,\ \sum_{i=1}^{10} y_i = 106,\ \sum_{i=1}^{10} x_i^2 = 970 \text{ e } \sum_{i=1}^{10} y_i^2 = 1.152.$$

Admitindo que, para ambos os grupos, modelos com distribuição Normal com mesma variância são adequados, o que pode ser concluído para um nível de significância $\alpha = 0{,}05$?

6. Para estudar o impacto de cenas violentas em desenhos animados, 5 crianças tiveram seus batimentos cardíacos medidos, antes e após assistirem a um desenho comercialmente veiculado por uma grande emissora de TV. Os dados são apresentados a seguir:

	Batimentos	
Criança	Antes	Após
1	96	114
2	102	112
3	108	112
4	89	83
5	85	99

Qual a conclusão, considerando-se um nível de significância $\alpha = 0{,}02$? Que suposições são necessárias?

7. Deseja-se comparar o consumo de refrigerantes entre alunos do ensino fundamental e do ensino médio. Uma amostra desses alunos foi coletada e solicitou-se que indicassem o número de latas de refrigerantes que consumiam por dia. Os resultados, para cada grupo, foram os seguintes:

$$\text{Ensino fundamental: } \sum_{i=1}^{20} x_i = 24,\ \sum_{i=1}^{20} x_i^2 = 610;$$

$$\text{Ensino médio: } \sum_{i=1}^{15} y_i = 34,\ \sum_{i=1}^{15} y_i^2 = 1.315.$$

Supondo normalidade, verifique se as variâncias populacionais são iguais (use $\alpha = 10\%$). Com base no resultado obtido, construa um teste de hipóteses para decidir, ao nível 1%, se os grupos têm o mesmo consumo médio.

8. O tempo de conexão à Internet tem sido modelado pela distribuição Normal com desvio padrão de 10 e 7 minutos para o público jovem e adulto, respectivamente. Amostras independentes desses dois públicos produziram:

$$\text{Adultos: } \sum_{i=1}^{6} x_i = 211{,}2;$$

$$\text{Jovens: } \sum_{i=1}^{10} y_i = 367.$$

Teste, usando $\alpha = 5\%$, se o tempo médio de conexão dos jovens é maior.

9. Deseja-se comparar a durabilidade de amortecedores fabricados pelas empresas A e B. A medida observada é o *índice de resistência* de cada peça testada em laboratório, que é assumido ter a mesma variabilidade nas duas empresas. Os resultados obtidos são apresentados a seguir.

Empresa A: 115, 123, 134, 120, 121 ;
Empresa B: 125, 126, 120, 130, 128 .

Fazendo as suposições necessárias, qual a conclusão ao nível 2%?

10. Num programa de diminuição da poluição sonora em cidades grandes, realizou-se uma campanha educativa durante 2 meses. A tabela abaixo apresenta os índices alcançados antes e após a campanha em dez pontos da cidade sorteados ao acaso.

	Pontos da Cidade									
	1	2	3	4	5	6	7	8	9	10
Antes	23	44	56	34	25	67	21	23	73	58
Depois	21	30	45	35	26	50	23	22	57	46

Verifique se a campanha surtiu efeito, ao nível de 4%. Indique as suposições que devem ser feitas.

11. Um experimento com cobaias consistia em comparar o desempenho de uma tarefa para dois métodos de aprendizagem. Cada cobaia teve seu desempenho medido, atribuindo-se uma nota de 0 a 10. Essa medição ocorreu imediatamente após cada aprendizagem. Os idealizadores do Método 2 alegam que seu método é mais eficiente e, portanto, deve produzir maior nota. De estudos anteriores, sabe-se que, após uma semana, o aprendizado de um método é esquecido e, portanto, fixou-se esse intervalo de tempo entre a aplicação dos métodos. Além disso, foi estabelecido que o modelo Normal é adequado para as variáveis envolvidas. As notas obtidas foram as seguintes:

Cobaia	1	2	3	4	5	6	7	8	9	10
Método 1	8	3	7	4	6	2	7	3	9	5
Método 2	8	4	8	3	7	3	8	3	8	6

Formule as hipóteses de interesse e faça o teste conveniente, considerando um nível de significância $\alpha = 5\%$.

12. Deseja-se comparar o tempo de recuperação pós-operatória para duas técnicas cirúrgicas. Pacientes operados, segundo cada uma das técnicas, foram selecionados aleatoriamente e seu tempo de recuperação, em dias, registrado. Todos os pacientes apresentavam o mesmo estado de saúde antes da cirurgia. Os dados obtidos são:

Técnica 1	4	4	5	6	6
Técnica 2	6	6	7	7	8

a. Verifique se as variâncias populacionais correspondentes às duas técnicas são iguais.

b. Verifique se a Técnica 1 é mais eficiente com relação ao tempo médio de recuperação. Use $\alpha = 5\%$.

13. Um casal está procurando um apartamento para alugar. Duas cidades, igualmente atrativas para o casal, estão sob consideração. A decisão sobre qual a cidade escolhida depende do custo do aluguel. Neste sentido, eles obtiveram o preço médio e a variabilidade dos preços em cada uma das cidades. Na primeira cidade, denominada cidade A, o preço médio e correspondente desvio padrão foram obtidos de uma pesquisa com 22 ofertas, fornecendo, respectivamente, R$ 455,00 e R$ 25,00. Na segunda cidade, B, foram escolhidas 30 ofertas que forneceram média de R$ 475,00 e desvio padrão de R$ 18,00. Supondo normalidade e mesma variância, pergunta-se:

a. As cidades são equivalentes, ao nível 5%? Se não, qual é melhor?

b. Apresente um intervalo de confiança com $\gamma = 0{,}95$ para a diferença de preços entre as cidades A e B.

14. Duas raças de cavalos estão sendo pesquisadas quanto à resistência a uma certa doença intestinal. O experimento consiste em ministrar uma dieta especial, durante 3 meses e verificar a ocorrência ou não da doença, nos 6 meses seguintes. Os dados observados foram os seguintes:

Raça	Tamanho da amostra	Ocorrências
A	58	6
B	85	10

Um veterinário afirma que a raça B tem maior propensão à doença. Você concorda? Use $\alpha = 6\%$.

15. Estão sendo testados dois aparelhos para estabelecer suas respectivas confiabilidades no diagnóstico de uma certa doença. Dados foram obtidos de

exames feitos em diversos pacientes, escolhidos ao acaso, que, após serem avaliados por um dos aparelhos, tiveram seus casos estudados em maior profundidade. Desse modo, foi possível quantificar o número de falsos positivos ou falsos negativos advindos do uso do aparelho. Em outras palavras, foi possível saber o número diagnosticado falsamente pelo aparelho como tendo ou não a doença. Seguem as informações obtidas:

Aparelho	Total	Positivos	Negativos	Falsos positivos	Falsos negativos
1	120	85	11	10	14
2	135	90	27	12	6

a. Teste se os dois aparelhos produzem a mesma proporção de diagnósticos falsos. Use $\alpha = 4\%$.

b. Dentre os que estão efetivamente doentes, isto é, os positivos e falsos negativos, teste se o aparelho 2 erra menos. Use $\alpha = 4\%$.

c. Com base nas decisões tomadas nos dois itens anteriores, que aparelho seria mais aconselhável utilizar?

16. Queremos comparar três hospitais, através da satisfação demonstrada por pacientes quanto ao atendimento, durante o período de internação. Para tanto, foram selecionados, aleatoriamente, pacientes com grau de enfermidade semelhante. Cada paciente preencheu um questionário e as respostas geraram índices variando de 0 a 100, indicando o grau de satisfação. Os resultados foram:

	Hospital		
	A	B	C
Tamanho da amostra	10	15	13
Média amostral	80,7	59,0	72,3
Variância amostral	113,3	101,4	106,5

a. Baseando-se nos dados apresentados, teste a igualdade das variâncias para os hospitais A e B. Use $\alpha = 0{,}10$.

b. Teste se as médias populacionais são iguais. Qual sua conclusão? Use $\alpha = 0{,}05$.

17. Pacientes resolveram processar a clínica de emagrecimento Linha Fina sob a alegação de que o tratamento empregado não contribui para a diminuição do peso. O advogado de defesa contratou um estatístico que selecionou, aleatoriamente, 10 prontuários que continham informação a respeito dos

pesos dos pacientes, tomados no início e no final do tratamento. Os dados obtidos foram (em kg):

	Número do paciente									
	1	2	3	4	5	6	7	8	9	10
Início	80	104	94	62	70	80	102	58	78	84
Final	78	95	87	60	71	82	94	65	78	80

a. Faça uma análise descritiva para os dados e obtenha uma conclusão preliminar.

b. Verifique se a conclusão do item anterior tem suporte estatístico. Formule as hipóteses adequadas e encontre a região crítica correspondente a $\alpha = 0{,}05$.

18. Uma linha de montagem utiliza robôs para a realização das tarefas necessárias para a montagem de um produto. Os técnicos acreditam que é necessário uma programação diferente para garantir a qualidade do produto final, mas suspeitam que o tempo necessário para completar o processo pode aumentar. Para verificar essa suspeita, 12 robôs foram selecionados e o tempo necessário para a montagem do produto foi medido, considerando-se a programação usual e a nova proposta. Os tempos observados (em minutos) para cada unidade foram medidos, produzindo a tabela a seguir.

Tipo de	Identificação do Robô											
Programação	1	2	3	4	5	6	7	8	9	10	11	12
Usual	80	90	93	92	75	92	72	87	90	86	78	97
Nova	100	85	90	102	90	99	97	95	100	94	89	98

a. Faça uma análise descritiva adequada a estes dados. O que pode ser dito, baseando-se nessa análise?

b. Existe diferença para os diferentes tipos de programação?

c. Construa um intervalo de confiança de confiança com $\gamma = 95\,\%$ para a diferença das médias populacionais dos tempos de montagem do produto.

19. O custo de manutenção de treminhões movidos a gasolina e a diesel são dados abaixo para duas amostras aleatórias de 10 treminhões de cada tipo. Os veículos considerados trafegam sob as mesmas condições em uma mesma área.

Combustível	Custos									
Gasolina	12,43	9,70	1,35	5,78	5,05	10,98	14,27	15,09	7,61	5,78
Diesel	12,26	5,13	3,32	14,72	4,17	12,72	8,89	9,95	2,94	5,06

a. Quais são as hipóteses necessárias para construir um intervalo de confiança, para a diferença das médias dos custos?

b. Verifique se as variâncias dos dois grupos são semelhantes.

c. Teste a igualdade de médias dos dois grupos, considerando $\alpha = 5\%$.

20. (Use o computador) Uma loja de departamentos está interessada em saber se existem diferenças entre as quantias médias faturadas, através de três formas de pagamento: dinheiro (D), cheque (C) e cartão de crédito (CC). Um levantamento das vendas (em milhares de reais), em um dado período de tempo, foi feito, produzindo os dados na tabela a seguir.

Formas de Pagamento		
D	C	CC
56,00	80,90	73,25
20,50	51,29	56,65
37,37	40,95	123,21
28,64	72,65	56,50
	132,47	37,29
	60,32	44,65
	60,00	40,64

a. Calcule algumas medidas descritivas (média, variância, etc.) e, baseado nelas, discuta se existem evidências de diferenças.

b. Assumindo que as variâncias são iguais para os três grupos, compare estatisticamente as médias populacionais para verificar se existem diferenças. Use $\alpha = 0{,}05$.

21. Quatro diferentes espécies de milho foram produzidas em laboratório. Deseja-se testar, a um nível de significância $\alpha = 5\%$, se existe diferença estatisticamente significativa entre as produtividades. Para tanto, foram montados 34 canteiros, plantando-se neles o mesmo número de sementes e garantindo-se a todos as mesmas condições de fertilidade, irrigação e exposição à luz solar. Após um período de tempo pré-especificado, a produção de cada canteiro, em kg, foi obtida. Os resultados observados

sofreram um tratamento inicial, e as seguintes informações foram disponibilizadas:

	Espécies			
i	1	2	3	4
n_i	9	10	7	8
$\overline{Y}_i$	90,56	86,40	95,71	83,63
S_i^2	13,28	14,27	13,23	16,55

O que pode ser concluído com base na tabela?

22. (Use o computador) O custo mensal de manutenção de determinado tipo de automóvel (excluindo-se combustível e trocas de óleo) está sendo analisado em função da idade do veículo. Nove automóveis fabricados em diferentes anos tiveram o custo averiguado. Os dados obtidos foram:

	Idade do veículo (anos)								
	1	2	3	4	5	6	7	8	9
Custo mensal (reais)	8	13	18	20	24	26	29	32	37

a. Faça um gráfico de dispersão e calcule o coeficiente de correlação. Comente o resultado.

b. Ajuste a reta de regressão pelo método de mínimos quadrados. Como você interpretaria o coeficiente β?

23. Verifique se é razoável considerar um modelo de regressão linear relacionando as notas de Inglês (Y) e Português (X), segundo os dados apresentados na tabela a seguir. Sua conclusão deve ser baseada no coeficiente de correlação e no ajuste da reta de regressão.

Disciplinas	Notas									
Inglês	5,5	3,5	7,0	2,5	8,5	6,5	6,0	4,0	0,5	5,0
Português	7,0	4,5	8,5	3,5	9,0	4,5	5,0	5,5	1,5	6,5

24. (Use o computador) Uma indústria submete seus novos operários a um teste de aptidão (X) e três meses depois mede a produtividade destes operários (Y). Os resultados estão na tabela a seguir

Operário	A	B	C	D	E	F
Aptidão (X)	22	25	15	19	22	18
Produtividade (Y)	45	37	25	40	33	30

a. Faça o gráfico de dispersão e calcule o coeficiente de correlação. Comente o resultado.

b. Ajuste a reta de regressão e trace-a no gráfico de dispersão. Qual a interpretação para os coeficientes α e β?

c. Verifique estatisticamente se a produtividade é influenciada pela aptidão. Use $\alpha = 0{,}05$.

d. Para um indivíduo com aptidão igual a 20, qual seria a produtividade esperada?

25. Um estudo pretende avaliar o efeito da obesidade na pressão sangüínea. Para tanto, foram avaliados os pesos para 6 indivíduos e construída a variável X representando a razão entre os pesos real e ideal. Estudos indicam que um modelo de regressão linear simples é adequado para essa situação. Os dados obtidos foram:

Indivíduo	1	2	3	4	5	6
Razão (X)	1,23	1,42	1,35	1,67	1,65	1,56
Pressão sistólica (Y)	129	130	133	139	136	134

a. Construa a variável auxiliar $d = x - \overline{x}$.

b. Ajuste a reta $y = \alpha + \beta\, d$.

c. Qual a interpretação para α na reta obtida em (b)?

d. Qual a pressão sistólica esperada para indivíduos com razão peso real/peso ideal igual a 1,25?

26. Estuda-se a relação linear entre duas variáveis X e Y. Uma amostra de 20 pares dessas variáveis forneceu os seguintes valores:

$$\sum_{i=1}^{20} x_i = 600,\ \sum_{i=1}^{20} y_i = 2.150,\ \sum_{i=1}^{20} x_i^2 = 18.662,\ \sum_{i=1}^{20} y_i^2 = 235.270,\ \sum_{i=1}^{20} x_i y_i = 65.921\,.$$

Determine a correlação e ajuste uma reta aos dados.

27. Um estudo foi conduzido para verificar se as pessoas estimam os próprios pesos corretamente. No experimento realizado, 15 pessoas foram selecionadas

ao acaso e a cada uma delas perguntou-se os pesos, que depois foram aferidos em balanças devidamente calibradas. Os resultados são apresentados a seguir.

	Indivíduo														
Peso	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Estimado	82	58	69	70	54	62	92	75	45	81	78	65	56	63	70
Real	83	57	73	76	55	60	98	74	44	82	76	67	54	60	71

O que pode ser concluído a partir dos dados?

28. A quantidade de chuva é um fator importante na produtividade agrícola. Para medir esse efeito, foram anotadas, para 8 regiões diferentes produtoras de soja, o índice pluviométrico e a produção do último ano.

Chuva (mm)	120	140	122	150	115	190	130	118
Produção (ton)	40	46	45	37	25	54	33	30

a. Faça um gráfico de dispersão e calcule o coeficiente de correlação. Comente o resultado.

b. Ajuste a reta de regressão. Como você interpretaria o coeficiente β?

c. Utilizando a reta ajustada no item (c), encontre a produção esperada para uma região com índice pluviométrico igual a 160 mm.

d. É adequado utilizar o modelo ajustado para calcular a produção em uma região cujo índice pluviométrico é igual a 30 mm? Comente.

29. (Use o computador) Para os dados do arquivo *areas.txt* (veja Exercício 25 do Capítulo 1), suponha que os apartamentos são classificados como de andar baixo, para a unidade situada entre os 1º ao 6º andares (inclusive); intermediário, se o apartamento se encontra do 7º ao 12º andar e de andar alto, se estiver situado acima do 12º andar. Suspeita-se que apartamentos de andares mais baixos podem não ter o tamanho especificado no memorial descritivo.

a. Construa histogramas para as medidas de área da sala para cada um dos grupos formados (andar baixo, intermediário e alto). Discuta se o modelo Normal é adequado para essa situação

b. Dependendo de cada uma das três categorias de andar criadas, construa um gráfico *box-plot* para as áreas da sala e discuta se a localização interfere com o tamanho da sala.

c. Usando um modelo de análise de variância, verifique se existem evidências estatísticas que dêem suporte à conclusão apresentada no item (b). Utilize um nível de significância $\alpha = 0{,}01$.

d. Repita os itens anteriores para a área total do apartamento, definida como a soma das áreas da sala, cozinha, banheiro e dormitório.

30. (Use o computador) Para este exercício será necessário utilizar o arquivo *cancer.txt*, cuja descrição é dada no Exercício 24 do Capítulo 1. Deseja-se verificar se conforme aumenta a idade, muda a concentração de nitrogênio na uréia.

a. Suponha que selecionamos apenas os pacientes que têm a doença (isto é, consideramos o grupo formado por pacientes cujo diagnóstico é falso-negativo ou positivo). Construa um gráfico de dispersão para idade e concentração de nitrogênio. O que pode ser dito?

b. Supondo que a variável dependente é a concentração de nitrogênio e que a covariável é a idade do paciente, calcule estimativas para α e β, em um modelo de regressão linear. Qual é a interpretação de β nesse caso?

c. Construa uma tabela ANOVA para verificar, ao nível de 5%, se existe evidência estatística de que a idade influencia na concentração de nitrogênio.

d. Considere, agora, os pacientes que não têm a doença (diagnóstico negativo ou falso-positivo). Construa um gráfico de dispersão para idade e concentração de nitrogênio. Compare com o gráfico obtido no item (a).

e. Calcule as estimativas para α e β nesse caso e interprete β. O que pode ser dito ao se comparar com os resultados do item (b)?

f. Verifique se a idade influencia a concentração de nitrogênio para os pacientes sem a doença. Considere $\alpha = 5\%$.

g. Com base nos itens anteriores, compare visualmente as retas ajustadas e as estimativas obtidas. Você diria que o efeito da idade, na concentração de nitrogênio, é um dado importante para discriminar entre pacientes com e sem a doença?

31. (Use o computador) Para este exercício será necessário utilizar o arquivo *aeusp.txt*, cuja descrição é dada no Exercício 26 do Capítulo 1.

a. Teste se a média da variável Itrab é a mesma nas sub-populações definidas pelo estado civil dos residentes.

b. Repita o item (a) com as sub-populações definidas pelo local de moradia.

Apêndice A

Tabelas

Distribuição Normal
Distribuição t - Student
Distribuição Qui-Quadrado
Distribuição Fisher-Snedecor (0,05)
Distribuição Fisher-Snedecor (0,95)

Distribuição Normal: Valores de p tais que $P(0 \leq Z \leq z_c) = p$

Segunda decimal de z_c

Parte inteira e primeira decimal de z_c	0	1	2	3	4	5	6	7	8	9	
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359	**0,0**
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753	**0,1**
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141	**0,2**
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517	**0,3**
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879	**0,4**
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224	**0,5**
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549	**0,6**
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852	**0,7**
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133	**0,8**
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389	**0,9**
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621	**1,0**
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830	**1,1**
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015	**1,2**
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177	**1,3**
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319	**1,4**
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441	**1,5**
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545	**1,6**
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633	**1,7**
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706	**1,8**
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767	**1,9**
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817	**2,0**
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857	**2,1**
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890	**2,2**
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916	**2,3**
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936	**2,4**
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952	**2,5**
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964	**2,6**
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974	**2,7**
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981	**2,8**
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986	**2,9**
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990	**3,0**
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993	**3,1**
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995	**3,2**
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997	**3,3**
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998	**3,4**
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	**3,5**
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	**3,6**
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	**3,7**
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	**3,8**
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	**3,9**

Distribuição *t*-Student : Valores t_c tais que $P(-t_c \le t \le t_c) = 1 - p$

Graus de liberdade	p->90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%	
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	15,894	31,821	63,656	318,289	636,578	1
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	4,849	6,965	9,925	22,328	31,600	2
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	3,482	4,541	5,841	10,214	12,924	3
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	2,999	3,747	4,604	7,173	8,610	4
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	2,757	3,365	4,032	5,894	6,869	5
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	2,612	3,143	3,707	5,208	5,959	6
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,517	2,998	3,499	4,785	5,408	7
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,449	2,896	3,355	4,501	5,041	8
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,398	2,821	3,250	4,297	4,781	9
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,359	2,764	3,169	4,144	4,587	10
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,328	2,718	3,106	4,025	4,437	11
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,303	2,681	3,055	3,930	4,318	12
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,282	2,650	3,012	3,852	4,221	13
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,264	2,624	2,977	3,787	4,140	14
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,249	2,602	2,947	3,733	4,073	15
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,235	2,583	2,921	3,686	4,015	16
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,224	2,567	2,898	3,646	3,965	17
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,214	2,552	2,878	3,610	3,922	18
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,205	2,539	2,861	3,579	3,883	19
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,197	2,528	2,845	3,552	3,850	20
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,189	2,518	2,831	3,527	3,819	21
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,183	2,508	2,819	3,505	3,792	22
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,177	2,500	2,807	3,485	3,768	23
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,172	2,492	2,797	3,467	3,745	24
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,167	2,485	2,787	3,450	3,725	25
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,162	2,479	2,779	3,435	3,707	26
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,158	2,473	2,771	3,421	3,689	27
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,154	2,467	2,763	3,408	3,674	28
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,150	2,462	2,756	3,396	3,660	29
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,147	2,457	2,750	3,385	3,646	30
35	0,127	0,255	0,388	0,529	0,682	0,852	1,052	1,306	1,690	2,030	2,133	2,438	2,724	3,340	3,591	35
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,123	2,423	2,704	3,307	3,551	40
50	0,126	0,255	0,388	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,109	2,403	2,678	3,261	3,496	50
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,099	2,390	2,660	3,232	3,460	60
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,076	2,358	2,617	3,160	3,373	120
∞	0,126	0,253	0,385	0,524	0,675	0,842	1,036	1,282	1,645	1,960	2,054	2,327	2,576	3,091	3,291	∞

Distribuição de Qui-quadrado: Valores de v_c tais que $P(\chi^2 \ge v_c) = p$

Graus de liberdade	p-> 99%	98%	97.5%	95%	90%	80%	70%	50%	30%	10%	5%	4%	2,5%	2%	1,0%	0,2%	0,1%	
1	0,000	0,001	0,001	0,004	0,016	0,064	0,148	0,455	1,074	2,706	3,841	4,218	5,024	5,412	6,635	9,549	10,827	1
2	0,020	0,040	0,051	0,103	0,211	0,446	0,713	1,386	2,408	4,605	5,991	6,438	7,378	7,824	9,210	12,429	13,815	2
3	0,115	0,185	0,216	0,352	0,584	1,005	1,424	2,366	3,665	6,251	7,815	8,311	9,348	9,837	11,345	14,796	16,266	3
4	0,30	0,43	0,48	0,71	1,06	1,65	2,19	3,36	4,88	7,78	9,49	10,03	11,14	11,67	13,28	16,92	18,47	4
5	0,55	0,75	0,83	1,15	1,61	2,34	3,00	4,35	6,06	9,24	11,07	11,64	12,83	13,39	15,09	18,91	20,51	5
6	0,87	1,13	1,24	1,64	2,20	3,07	3,83	5,35	7,23	10,64	12,59	13,20	14,45	15,03	16,81	20,79	22,46	6
7	1,24	1,56	1,69	2,17	2,83	3,82	4,67	6,35	8,38	12,02	14,07	14,70	16,01	16,62	18,48	22,60	24,32	7
8	1,65	2,03	2,18	2,73	3,49	4,59	5,53	7,34	9,52	13,36	15,51	16,17	17,53	18,17	20,09	24,35	26,12	8
9	2,09	2,53	2,70	3,33	4,17	5,38	6,39	8,34	10,66	14,68	16,92	17,61	19,02	19,68	21,67	26,06	27,88	9
10	2,56	3,06	3,25	3,94	4,87	6,18	7,27	9,34	11,78	15,99	18,31	19,02	20,48	21,16	23,21	27,72	29,59	10
11	3,05	3,61	3,82	4,57	5,58	6,99	8,15	10,34	12,90	17,28	19,68	20,41	21,92	22,62	24,73	29,35	31,26	11
12	3,57	4,18	4,40	5,23	6,30	7,81	9,03	11,34	14,01	18,55	21,03	21,79	23,34	24,05	26,22	30,96	32,91	12
13	4,11	4,77	5,01	5,89	7,04	8,63	9,93	12,34	15,12	19,81	22,36	23,14	24,74	25,47	27,69	32,54	34,53	13
14	4,66	5,37	5,63	6,57	7,79	9,47	10,82	13,34	16,22	21,06	23,68	24,49	26,12	26,87	29,14	34,09	36,12	14
15	5,23	5,98	6,26	7,26	8,55	10,31	11,72	14,34	17,32	22,31	25,00	25,82	27,49	28,26	30,58	35,63	37,70	15
16	5,81	6,61	6,91	7,96	9,31	11,15	12,62	15,34	18,42	23,54	26,30	27,14	28,85	29,63	32,00	37,15	39,25	16
17	6,41	7,25	7,56	8,67	10,09	12,00	13,53	16,34	19,51	24,77	27,59	28,44	30,19	31,00	33,41	38,65	40,79	17
18	7,01	7,91	8,23	9,39	10,86	12,86	14,44	17,34	20,60	25,99	28,87	29,75	31,53	32,35	34,81	40,14	42,31	18
19	7,63	8,57	8,91	10,12	11,65	13,72	15,35	18,34	21,69	27,20	30,14	31,04	32,85	33,69	36,19	41,61	43,82	19
20	8,26	9,24	9,59	10,85	12,44	14,58	16,27	19,34	22,77	28,41	31,41	32,32	34,17	35,02	37,57	43,07	45,31	20
21	8,90	9,91	10,28	11,59	13,24	15,44	17,18	20,34	23,86	29,62	32,67	33,60	35,48	36,34	38,93	44,52	46,80	21
22	9,54	10,60	10,98	12,34	14,04	16,31	18,10	21,34	24,94	30,81	33,92	34,87	36,78	37,66	40,29	45,96	48,27	22
23	10,20	11,29	11,69	13,09	14,85	17,19	19,02	22,34	26,02	32,01	35,17	36,13	38,08	38,97	41,64	47,39	49,73	23
24	10,86	11,99	12,40	13,85	15,66	18,06	19,94	23,34	27,10	33,20	36,42	37,39	39,36	40,27	42,98	48,81	51,18	24
25	11,52	12,70	13,12	14,61	16,47	18,94	20,87	24,34	28,17	34,38	37,65	38,64	40,65	41,57	44,31	50,22	52,62	25
26	12,20	13,41	13,84	15,38	17,29	19,82	21,79	25,34	29,25	35,56	38,89	39,89	41,92	42,86	45,64	51,63	54,05	26
27	12,88	14,13	14,57	16,15	18,11	20,70	22,72	26,34	30,32	36,74	40,11	41,13	43,19	44,14	46,96	53,02	55,48	27
28	13,56	14,85	15,31	16,93	18,94	21,59	23,65	27,34	31,39	37,92	41,34	42,37	44,46	45,42	48,28	54,41	56,89	28
29	14,26	15,57	16,05	17,71	19,77	22,48	24,58	28,34	32,46	39,09	42,56	43,60	45,72	46,69	49,59	55,79	58,30	29
30	14,95	16,31	16,79	18,49	20,60	23,36	25,51	29,34	33,53	40,26	43,77	44,83	46,98	47,96	50,89	57,17	59,70	30

Distribuição de Fisher-Snedecor - Valores de f_c tais que $P(F \geq f_c) = 0{,}05$

Graus de liberdade do numerador

Graus de liberdade do denominador

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	60	120	∞	
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,0	243,9	244,7	245,4	245,9	246,5	246,9	247,3	247,7	248,0	249,3	250,1	252,2	253,3	254,3	1
2	18,513	19,000	19,164	19,247	19,296	19,329	19,353	19,371	19,385	19,396	19,405	19,412	19,419	19,424	19,429	19,433	19,437	19,440	19,443	19,446	19,456	19,463	19,479	19,487	19,496	2
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785	8,763	8,745	8,729	8,715	8,703	8,692	8,683	8,675	8,667	8,660	8,634	8,617	8,572	8,549	8,526	3
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,936	5,912	5,891	5,873	5,858	5,844	5,832	5,821	5,811	5,803	5,769	5,746	5,688	5,658	5,628	4
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,704	4,678	4,655	4,636	4,619	4,604	4,590	4,579	4,568	4,558	4,521	4,496	4,431	4,398	4,365	5
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,027	4,000	3,976	3,956	3,938	3,922	3,908	3,896	3,884	3,874	3,835	3,808	3,740	3,705	3,669	6
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,603	3,575	3,550	3,529	3,511	3,494	3,480	3,467	3,455	3,445	3,404	3,376	3,304	3,267	3,230	7
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388	3,347	3,313	3,284	3,259	3,237	3,218	3,202	3,187	3,173	3,161	3,150	3,108	3,079	3,005	2,967	2,928	8
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,102	3,073	3,048	3,025	3,006	2,989	2,974	2,960	2,948	2,936	2,893	2,864	2,787	2,748	2,707	9
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,943	2,913	2,887	2,865	2,845	2,828	2,812	2,798	2,785	2,774	2,730	2,700	2,621	2,580	2,538	10
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,818	2,788	2,761	2,739	2,719	2,701	2,685	2,671	2,658	2,646	2,601	2,570	2,490	2,448	2,404	11
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,717	2,687	2,660	2,637	2,617	2,599	2,583	2,568	2,555	2,544	2,498	2,466	2,384	2,341	2,296	12
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,635	2,604	2,577	2,554	2,533	2,515	2,499	2,484	2,471	2,459	2,412	2,380	2,297	2,252	2,206	13
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,565	2,534	2,507	2,484	2,463	2,445	2,428	2,413	2,400	2,388	2,341	2,308	2,223	2,178	2,131	14
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,507	2,475	2,448	2,424	2,403	2,385	2,368	2,353	2,340	2,328	2,280	2,247	2,160	2,114	2,066	15
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,456	2,425	2,397	2,373	2,352	2,333	2,317	2,302	2,288	2,276	2,227	2,194	2,106	2,059	2,010	16
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,413	2,381	2,353	2,329	2,308	2,289	2,272	2,257	2,243	2,230	2,181	2,148	2,058	2,011	1,960	17
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,374	2,342	2,314	2,290	2,269	2,250	2,233	2,217	2,203	2,191	2,141	2,107	2,017	1,968	1,917	18
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378	2,340	2,308	2,280	2,256	2,234	2,215	2,198	2,182	2,168	2,155	2,106	2,071	1,980	1,930	1,878	19
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,310	2,278	2,250	2,225	2,203	2,184	2,167	2,151	2,137	2,124	2,074	2,039	1,946	1,896	1,843	20
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420	2,366	2,321	2,283	2,250	2,222	2,197	2,176	2,156	2,139	2,123	2,109	2,096	2,045	2,010	1,916	1,866	1,812	21
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,342	2,297	2,259	2,226	2,198	2,173	2,151	2,131	2,114	2,098	2,084	2,071	2,020	1,984	1,889	1,838	1,783	22
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,320	2,275	2,236	2,204	2,175	2,150	2,128	2,109	2,091	2,075	2,061	2,048	1,996	1,961	1,865	1,813	1,757	23
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,300	2,255	2,216	2,183	2,155	2,130	2,108	2,088	2,070	2,054	2,040	2,027	1,975	1,939	1,842	1,790	1,733	24
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282	2,236	2,198	2,165	2,136	2,111	2,089	2,069	2,051	2,035	2,021	2,007	1,955	1,919	1,822	1,768	1,711	25
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,265	2,220	2,181	2,148	2,119	2,094	2,072	2,052	2,034	2,018	2,003	1,990	1,938	1,901	1,803	1,749	1,691	26
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305	2,250	2,204	2,166	2,132	2,103	2,078	2,056	2,036	2,018	2,002	1,987	1,974	1,921	1,884	1,785	1,731	1,672	27
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291	2,236	2,190	2,151	2,118	2,089	2,064	2,041	2,021	2,003	1,987	1,972	1,959	1,906	1,869	1,769	1,714	1,654	28
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,223	2,177	2,138	2,104	2,075	2,050	2,027	2,007	1,989	1,973	1,958	1,945	1,891	1,854	1,754	1,698	1,638	29
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,126	2,092	2,063	2,037	2,015	1,995	1,976	1,960	1,945	1,932	1,878	1,841	1,740	1,683	1,622	30
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,952	1,917	1,887	1,860	1,836	1,815	1,796	1,778	1,763	1,748	1,690	1,649	1,534	1,467	1,389	60
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910	1,869	1,834	1,803	1,775	1,750	1,728	1,709	1,690	1,674	1,659	1,598	1,554	1,429	1,352	1,254	120
∞	3,842	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880	1,831	1,789	1,752	1,720	1,692	1,666	1,644	1,623	1,604	1,587	1,571	1,507	1,459	1,318	1,221	1,000	∞

Distribuição de Fisher-Snedecor - Valores de f_c tais que $P(F \geq f_c) = 0{,}95$

Graus de liberdade do numerador

Graus de liberdade do denominador

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	60	120	∞	
1	,006	,054	,099	,130	,151	,167	,179	,188	,195	,201	,206	,211	,214	,217	,220	,223	,225	,227	,228	,230	,236	,240	,250	,255	,260	1
2	,005	,053	,105	,144	,173	,194	,211	,224	,235	,244	,251	,257	,263	,267	,272	,275	,278	,281	,284	,286	,295	,302	,317	,326	,334	2
3	,005	,052	,108	,152	,185	,210	,230	,246	,259	,270	,279	,287	,293	,299	,304	,309	,313	,316	,320	,323	,334	,342	,363	,373	,384	3
4	,004	,052	,110	,157	,193	,221	,243	,261	,275	,288	,298	,307	,315	,321	,327	,333	,337	,342	,345	,349	,362	,372	,396	,409	,422	4
5	,004	,052	,111	,160	,198	,228	,252	,271	,287	,301	,312	,322	,331	,338	,345	,351	,356	,361	,365	,369	,384	,395	,422	,437	,452	5
6	,004	,052	,112	,162	,202	,233	,259	,279	,296	,311	,323	,334	,343	,351	,358	,365	,371	,376	,380	,385	,402	,413	,444	,460	,477	6
7	,004	,052	,113	,164	,205	,238	,264	,286	,304	,319	,332	,343	,353	,362	,369	,376	,383	,388	,393	,398	,416	,428	,462	,479	,498	7
8	,004	,052	,113	,166	,208	,241	,268	,291	,310	,326	,339	,351	,361	,371	,379	,386	,392	,398	,404	,409	,428	,441	,477	,496	,516	8
9	,004	,052	,113	,167	,210	,244	,272	,295	,315	,331	,345	,358	,368	,378	,386	,394	,401	,407	,413	,418	,438	,452	,490	,511	,532	9
10	,004	,052	,114	,168	,211	,246	,275	,299	,319	,336	,350	,363	,374	,384	,393	,401	,408	,415	,421	,426	,447	,462	,502	,523	,546	10
11	,004	,052	,114	,168	,213	,248	,278	,302	,322	,340	,355	,368	,380	,390	,399	,407	,414	,421	,427	,433	,455	,470	,512	,535	,559	11
12	,004	,052	,114	,169	,214	,250	,280	,305	,325	,343	,359	,372	,384	,395	,404	,412	,420	,427	,433	,439	,462	,478	,522	,545	,571	12
13	,004	,051	,115	,170	,215	,251	,282	,307	,328	,346	,362	,376	,388	,399	,408	,417	,425	,432	,439	,445	,468	,485	,530	,555	,581	13
14	,004	,051	,115	,170	,216	,253	,283	,309	,331	,349	,365	,379	,392	,403	,412	,421	,429	,437	,443	,449	,474	,491	,538	,563	,591	14
15	,004	,051	,115	,171	,217	,254	,285	,311	,333	,351	,368	,382	,395	,406	,416	,425	,433	,441	,448	,454	,479	,496	,545	,571	,600	15
16	,004	,051	,115	,171	,217	,255	,286	,312	,335	,354	,370	,385	,398	,409	,419	,429	,437	,445	,451	,458	,483	,501	,551	,579	,608	16
17	,004	,051	,115	,171	,218	,256	,287	,314	,336	,356	,372	,387	,400	,412	,422	,432	,440	,448	,455	,462	,487	,506	,557	,585	,616	17
18	,004	,051	,115	,172	,218	,257	,288	,315	,338	,357	,374	,389	,403	,414	,425	,434	,443	,451	,458	,465	,491	,510	,562	,592	,623	18
19	,004	,051	,115	,172	,219	,257	,289	,316	,339	,359	,376	,391	,405	,417	,427	,437	,446	,454	,461	,468	,495	,514	,567	,597	,630	19
20	,004	,051	,115	,172	,219	,258	,290	,317	,341	,360	,378	,393	,407	,419	,430	,439	,448	,456	,464	,471	,498	,518	,572	,603	,637	20
21	,004	,051	,116	,173	,220	,259	,291	,318	,342	,362	,379	,395	,409	,421	,432	,442	,451	,459	,466	,473	,501	,521	,577	,608	,643	21
22	,004	,051	,116	,173	,220	,259	,292	,319	,343	,363	,381	,396	,410	,423	,434	,444	,453	,461	,469	,476	,504	,524	,581	,613	,648	22
23	,004	,051	,116	,173	,221	,260	,293	,320	,344	,364	,382	,398	,412	,424	,435	,446	,455	,463	,471	,478	,507	,527	,585	,617	,654	23
24	,004	,051	,116	,173	,221	,260	,293	,321	,345	,365	,383	,399	,413	,426	,437	,447	,457	,465	,473	,480	,509	,530	,588	,622	,659	24
25	,004	,051	,116	,173	,221	,261	,294	,322	,346	,366	,384	,400	,415	,427	,439	,449	,458	,467	,475	,482	,511	,532	,592	,626	,664	25
26	,004	,051	,116	,174	,221	,261	,294	,322	,346	,367	,385	,402	,416	,429	,440	,451	,460	,469	,477	,484	,514	,535	,595	,630	,669	26
27	,004	,051	,116	,174	,222	,262	,295	,323	,347	,368	,386	,403	,417	,430	,441	,452	,462	,470	,478	,486	,516	,537	,598	,633	,673	27
28	,004	,051	,116	,174	,222	,262	,295	,324	,348	,369	,387	,404	,418	,431	,443	,453	,463	,472	,480	,487	,518	,539	,601	,637	,677	28
29	,004	,051	,116	,174	,222	,262	,296	,324	,349	,370	,388	,405	,419	,432	,444	,455	,464	,473	,481	,489	,519	,541	,604	,640	,681	29
30	,004	,051	,116	,174	,222	,263	,296	,325	,349	,370	,389	,405	,420	,433	,445	,456	,466	,475	,483	,490	,521	,543	,606	,643	,685	30
60	,004	,051	,117	,176	,226	,267	,303	,333	,359	,382	,402	,419	,435	,450	,463	,475	,486	,496	,505	,514	,549	,575	,652	,700	,759	60
120	,004	,051	,117	,177	,227	,270	,306	,337	,364	,388	,408	,427	,444	,459	,473	,486	,497	,508	,518	,527	,565	,594	,682	,740	,819	120
∞	,004	,051	,117	,178	,229	,273	,310	,342	,369	,394	,416	,436	,453	,469	,484	,498	,510	,522	,532	,543	,584	,616	,720	,798	,997	∞

Apêndice B

Respostas dos Exercícios

Observações:

1. Nos exercícios de seção a resposta será, na maioria das vezes, acompanhada de indicações da resolução.
2. Para os exercícios de fim de capítulo serão apresentadas as respostas para os exercícios ímpares.
3. Os exercícios de computação e de demonstração não terão suas respostas apresentadas.
4. Pequenas diferenças em algumas respostas poderão refletir diferentes aproximações e casas decimais utilizadas.
5. Para não tornar muito extenso esse apêndice, os gráficos solicitados foram omitidos na apresentação das respostas.